IOWA STATE UNIVERSITY
Digital Repository

Retrospective Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2003

Spliced alignment and its application in Arabidopsis thaliana

Wei Zhu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

 Part of the Genetics Commons

Recommended Citation
Zhu, Wei, "Spliced alignment and its application in Arabidopsis thaliana " (2003). *Retrospective Theses and Dissertations*. 1411.
https://lib.dr.iastate.edu/rtd/1411

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Spliced alignment and its application in *Arabidopsis thaliana*

by

Wei Zhu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Volker Brendel, Co-major Professor
Srinivas Aluru, Co-major Professor
Thomas Peterson
Gavin J.P. Naylor
Patrick S. Schnable

Iowa State University

Ames, Iowa

2003

UMI Number: 3085964

UMI®

UMI Microform 3085964
Copyright 2003 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Wei Zhu

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Major Program

For my family and my teachers

TABLE OF CONTENTS

ABSTRACT

This thesis describes the development and biological applications of GeneSeqer, which is a homology-based gene prediction program by means of spliced alignment. Additionally, a program named MyGV was written in JAVA as a browser to visualize the output of GeneSeqer. In order to test and demonstrate the performance, GeneSeqer was utilized to map 176,915 *Arabidopsis* EST sequences on the whole genome of *Arabidopsis thaliana*, which consists of five chromosomes, with about 117 million base pairs in total. All results were parsed and imported into a MySQL database. Information that was inferred from the *Arabidopsis* spliced alignment results may serve as valuable resource for a number of projects of special scientific interest, such as alternative splicing, non-canonical splice sites, mini-exons, etc. We also built AtGDB (*Arabidopsis thaliana* Genome DataBase, http://www.plantgdb.org/AtGDB/) to interactively browse EST spliced alignments and GenBank annotations for the *Arabidopsis* genome. Moreover, as one application of the *Arabidopsis* EST mapping data, U12-type introns were identified from the transcript-confirmed introns in the *Arabidopsis* genome, and the characteristics of these minor class introns were further explored.

CHAPTER I. GENERAL INTRODUCTION

Introduction

The growing number of completed genome sequencing projects demands high-throughput genome annotations. Genome annotation originally relied primarily upon *ab initio* gene prediction methods. To date, the accuracy of the leading *ab initio* gene prediction programs has reached slightly more than 90% at the nucleotide level, and on average 80% of exons can be precisely located (Burset and Guigo, 1996; Pavy et al., 1999; Guigo et al., 2000; Reese et al., 2000a). Provided that the average number of exons per gene is five, however, only about 33% of genes could be exactly identified. In addition, there is high rate of false positives when applying *ab initio* gene finders on large genomic data (Guigo et al., 2000). Alternative splicing, which brings up the unexpected genome complexity (Mironov et al., 1999; Brett et al., 2002; Modrek and Lee, 2002), makes genome annotation a more difficult task to be handled by *ab initio* gene prediction programs alone, because most of those programs make only one optimal prediction for each gene locus. On the other hand, spliced alignment can reveal gene structure more accurately by aligning the transcription or translation products with the genomic sequence, which is also called similarity-based gene prediction. The widespread application of similarity-based gene prediction in automated genome annotation is also inspired by the exponential accumulation of sequence data. A number of tools have been developed to address the issue and have already been utilized in genome annotation (Mathe et al., 2002). To achieve high speed, most of those programs have adopted heuristic method for aligning and/or utilizing the simple GT/AG rule for the recognition of the exon-intron junction, which lead to poor performance in some cases, including low sequence similarity, non-canonical splice sites, or mini-exons (Haas et al., 2002). To address this problem, we developed the GeneSeqer program to generate spliced alignments with high accuracy. The design, implementation and application of GeneSeqer are described in this thesis.

Thesis Organization

This thesis adopts a journal paper format such that each paper or manuscript will appear as a separate chapter. In addition, a general introduction is given in this, the first chapter, and an overall conclusion is included as the last chapter of the thesis. Chapter 2 covers the algorithm and implementation of GeneSeqer, and its performance in plant genomes. Chapter 3 introduces MyGV, which enables users to browse the GeneSeqer output visually. It also allows the user to compare alignments with GenBank annotations and predictions from different gene finders, such as GENSCAN (Burge and Karlin, 1997), GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi) and others. This paper was published as an 'Application Note' in *Bioinformatics* (Zhu and Brendel, 2002), and a more detailed information is available at the download page of MyGV (http://bioinformatics.iastate.edu/bioinformatics2go/MyGV/). Chapter 4 illustrates a 'one gene at a time' method of combining *ab initio* gene predictions with similarity-based prediction methods (Brendel and Zhu, 2002). Chapter 5 describes the application of GeneSeqer on the genome level, as explored in *Arabidopsis thaliana* by mapping all *Arabidopsis* ESTs onto the *A. thaliana* genome (Zhu et al., 2003). As an application of the *Arabidopsis* EST mapping data from the previous study, Chapter 6 concentrates on the identification and characterization of U12-type introns in the *Arabidopsis* genome. Finally, Chapter 7 discusses features and existing problems of GeneSeqer and the interesting discoveries from the GeneSeqer application in *Arabidopsis*.

Literature Review

Characteristics and prediction of pre-mRNA splicing signals

Eukaryotic protein coding genes differ from those of prokaryotes in that the structural genes are 'in pieces', that is, the expressed sequences (exons) are interspersed with non-coding intervening sequences (introns). On the basis of primary sequence, secondary structure and splicing mechanisms, introns can be grouped into the following categories: spliceosome introns (reviewed by Burge et al., 1999), self-splicing introns (Group I, II and III), and archaeal and nuclear tRNA introns (Lykke-Andersen et al., 1997). The spliceosome introns, which are the most abundant introns in the nuclear

genome, are spliced via a two-step transesterification reaction by the spliceosome which is a large complex consisting of U-rich small nuclear RNAs (snRNAs) and numerous protein factors (reviewed by Burge et al., 1999). During the splicing process, the entire structural gene is initially transcribed (including introns) to form the pre-mRNA, and then the introns are removed and the flanking exons are merged to form the mature mRNA after capping and polyadenylation. The spliceosome introns are also called pre-mRNA introns or nuclear introns.

Donor site (also termed 5' splice site, or 5'ss), branchpoint sequence (BPS), and acceptor site (also termed 3' splice site, or 3'ss) are three major cis-acting sequence elements functioning in the pre-mRNA intron splicing process, with common consensus sequences /GTAYGU, CURAY, and YAG/, respectively (where / denotes the exon-intron junction and the branchsite adenosine is underlined). Those signal elements have similar patterns among vertebrates, yeast, and plants, except that the BPS exhibits a more conserved motif in yeast than the others. The motifs in 5'ss and BPS play an important role in intron recognition by the conventional (U2-dependent) spliceosome pathway, which is mediated by the base pairing between the 5'ss and U1 or U6 snRNAs and the base pairing between the BPS and U2 snRNA. However, there is another class of low abundance introns with unusual splice signal motifs that are spliced by the minor (U12-dependent) spliceosome (reviewed by Burge et al., 1999). It is noted that more than 98% of nuclear introns are U2-type GT-AG introns, following the so-called GT-AG rule (Burset et al., 2001; Clark and Thanaraj, 2002). The GT-AG introns are also referred to as conventional introns or canonical introns. Among the non-canonical introns, the majority are GC-AG introns or AT-AC introns, which are generally processed by the major and minor spliceosome pathways, respectively (Aebi et al., 1987; Tarn and Steitz, 1996a; Tarn and Steitz, 1996b). Because of the low abundance and unusual conserved features of the splice signals, U12-dependent intron prediction is addressed by particular methods not further discussed here (for large scale computational surveys of U12-dependent introns see Burge et al., 1998; Levine and Durbin, 2001; also see Chapter 6) . For simplicity, the term intron refers to pre-mRNA intron spliced by the conventional spliceosome in the following.

Besides the three common splice signals, the presence of a polypyrimidine tract between the BPS and 3'ss is prominent in vertebrate introns and some yeast introns. Plant introns have instead a strong compositional bias for UA-rich or U-rich sequences, which is essential for plant intron excision (Goodall and Filipowicz, 1989; Carle-Urioste, et al., 1997; Ko, et al., 1998). Differences in intron processing also exist between monocots and dicots among higher plants (Goodall and Filipowicz, 1991).

Other elements such as the exonic splicing enhancer or silencer also promote/suppress splicing efficiency (Hastings and Krainer, 2001). Nevertheless, the selection of the intron boundaries is largely dependent on the splice site sequences. Correspondingly, a number of splice site prediction programs attempt to recognize splice signals based on the local sequence by means of Markov model, maximal dependence decomposition method, artificial neural network, logitlinear model or other methods (reviewed by Mathe et al., 2002; and references therein). Hebsgaard et al. (1996) also explored ways of improving the prediction accuracy by incorporating global sequence information. However, the accuracy of the prediction is limited by the fact that the sequence elements alone are not sufficient to determine the exon-intron junction and many trans-acting elements also play an important role in the selection of the splice signals (Black, 2003). Nevertheless, splice site prediction is an important component in spliced alignment, genomic comparison and *ab initio* gene prediction.

Spliced alignment

Spliced alignment is the alignment of ESTs/cDNA or proteins with genomic sequences utilizing a specific gap penalty that permits long gaps corresponding to the intron locations (recently review by Mathe et al., 2002). The existing spliced alignment programs can be roughly categorized into four groups. One group uses full dynamic programming to find the optimal spliced alignment, including Procrustes (Gelfand et al., 1996), GeneWise (Birney and Durbin, 1997), est2gen (Birney and Durbin, 1997), EST_GENOME (Mott, 1997), and SplicePredictor (Usuka and Brendel, 2000). The dynamic programming finds the optimal solution at the cost of significant running time and computer resources. The second group utilizes a blast-like heuristic method that determines near-perfect matching regions (High-scoring Segment Pairs, HSPs) corresponding to exons and only applies dynamic programming in finding the highest scored chain of the segment pairs and filling the dangling region around exon-intron junctions, as sim4 (Florea et al., 1998) and Spidey (Wheelan et al., 2001). The heuristic method reduces the running time dramatically, for example, sim4 was reported 300 times faster than EST_GENOME (Florea et al., 1998). However, their accuracy may also drop when the sequence similarity is low. Additionally, all programs listed above rely on BLAST (Altschul et al., 1997) or other external search engines to identify the genomic location which matches the sequence evidence. To simplify the data processing and reduce the genomic localization time, another group of programs were developed with built-in search engines to determine the hit region in the genomic sequence, while maintaining the full dynamic programming for accuracy as in the first group, as AAT (Huang et al., 1997) and GeneSeqer (Usuka et al., 2000). To date, the new

generation of spliced alignment programs was targeted to align millions of EST/cDNAs with the human genome in a reasonable time, such as BLAT (Kent, 2002) and SQUALL (Ogasawara and Morishita, 2002). To archive the goal, the whole genome is typically indexed to facilitate the genomic localization and only the "best hit" (a perfect or near-perfect match) is aligned for each evidence sequence. Consequently, the speed is further improved, such that BLAT is 600 times faster than sim4 according to its benchmark (Kent, 2002). In practical testing, BLAT mapped 3 million ESTs on the $3x10^9$ bases of human DNA in about 10 days in a single processor (Kent, 2002), and SQUALL can finish the same task in less than 42 hours (Ogasawara and Morishita, 2002).

Genome annotation

Currently, alternative splicing is found to play an important role in human functional genomics (Brett et al., 2002), and the associated research also suggests that ESTs and EST spliced alignments are an indispensable way of unearthing alternative splicing (Modrek and Lee, 2002; Zhu and Brendel, 2002). The role of spliced alignments is evolving significantly as more experimental data become available.

Besides spliced alignments, genomic comparison emerges as a new data resource for the identification of gene structures. Functional sites may be more conserved than "junk" DNA in the genome during evolution (Hardison et al., 1997; Wasserman et al., 2000). Gene order in related-species may also be conserved, for example, large syntenic regions or collinearity has been identified between human and mouse (Mural et al., 2002; Xuan et al., 2003; Guigo et al., 2003), and *Brassica oleracea* and *Arabidopsis thaliana* (Lan et al., 2000). Furthermore, micro-synteny may exist in remotely-related species (Chen et al., 1997; Liu et al., 2001; Salse et al., 2002), and the synteny of some core orthologous genes might even be conserved across eukaryotes (Trachtulec and Forejt, 2001). Therefore, genomic comparison among syntenic regions across the species may not only reveal gene organization, but also help to locate regulatory elements. The sequencing of the mouse genome is nearly completed, which makes this technique more attractive. Dozens of programs have come out in the last a few years (Bafna and Huson, 2000; Novichkov et al., 2001), and others were reviewed by Miller (2001) and recently by Mathe et al. (2002).

Both spliced alignment and genomic comparison are categorized into similarity-based gene prediction methods, also called extrinsic approaches. In contrast to similarity-based gene prediction which is based on the homologous sequence evidence, *ab initio* gene prediction uses statistical and computational methods to build signals and content sensors to identify functional elements related

with gene structures such as promoters, splice sites, exons, introns, and translation initiation and termination sites. Most of *ab initio* gene finders are composed of several different specific sensors integrated by either dynamic programming or hidden Markov model (Pavy et al., 1999; Pavy et al., 1999; Guigo et al., 2000). Such *ab initio* gene prediction programs are generally very fast and require little space, and some of them can reach remarkable accuracy at the nucleotide level but the accuracy at the gene level is still less than 50%. Moreover, most of gene finders can not handle complicated gene structures and non-conventional biological signals, including: 1) alternative splicing; 2) nested/overlapped genes; 3) very long introns; 4) very short exons; 5) non-canonical introns; 6) frame-shift errors; 7) split start codons (that is, the start codon is split by an intron in the genomic sequence); 8) introns in non-coding regions.

The current trend is to integrate two complementary groups of gene identification methods in order to further improve the accuracy of gene prediction (recently reviewed by Mathe, et al., 2002). A number of *ab initio* gene prediction programs were modified to adopt homologous sequence information, such as GRAIL (Xu and Uberbacher, 1996), GenomeScan (Yeh et al., 2001), TwinScan (Korf et al., 2001), GeneId+ (Parra et al., 2000), GenieEST/GenieESTHOM (Reese et al., 2000b), FgeneSH+ (Salamov and Solovyev, 2000). This combination not only occurs within the sole program, but also among different programs, for example, GeneScope (Murakami and Takagi, 1998) and GeneMachine (Makalowska et al., 2001). Moreover, automated genome annotation can also be regarded as an amalgamation of different annotation tools at a higher level, such as RiceGAAS (Sakata et al., 2002), Ensembl (Hubbard et al., 2002), the automated mouse genome annotation (Mural et al., 2002), and UCSC genome browser (Kent et al., 2002; Karolchik et al., 2003). The assessment of such hybrid methods suggests that the accuracy of gene prediction is significantly improved (Salamov and Solovyev, 2000; Reese et al., 2000b; Yeh et al., 2001).

References

Aebi, M., Hornig, H., and Weissmann, C. (1987). 5' cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5' splice region, not by the conserved 5' GU. Cell *50*, 237-46.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-402.

Bafna, V. and Huson, D.H. (2000). The conserved exon method for gene finding. Proc Int Conf Intell Syst Mol Biol *8*, 3-12.

Birney, E. and Durbin, R. (1997). Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. Proc Int Conf Intell Syst Mol Biol *5*, 56-64.

Black, D.L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu Rev Biochem *72*, 291-336

Brendel, V. and Zhu, W. (2002). Computational modeling of gene structure in *Arabidopsis thaliana*. Plant Mol Biol *48*, 49-58.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat Genet *30*, 29-30.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Burge, C.B., Padgett, R.A., and Sharp, P.A. (1998). Evolutionary fates and origins of U12-type introns. Mol Cell *2*, 773-85.

Burge, C.B., Tuschl, T., and Sharp, P.A. (1999). Splicing of precursors to mRNAs by the spliceosome. In The RNA world II. R.F. Gestland, T. Cech, and J.F. Atkins, eds. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. p.525-560.

Burset, M. and Guigo, R. (1996). Evaluation of gene structure prediction programs. Genomics *34*, 353-67.

Burset, M., Seledtsov, I.A., and Solovyev, V.V. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res *29*, 255-9.

Chen, M., SanMiguel, P., de Oliveira, A.C., Woo, S.S., Zhang, H., Wing, R.A., and Bennetzen, J.L. (1997). Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. Proc Natl Acad Sci U S A *94*, 3431-5.

Carle-Urioste, J.C., Brendel, V., and Walbot, V. (1997). A combinatorial role for exon, intron and splice site sequences in splicing in maize. Plant J *11*, 1253-63.

Clark, F. and Thanaraj, T.A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet *11*, 451-64.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res *8*, 967-74.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. (1996). Gene recognition via spliced sequence alignment. Proc Natl Acad Sci U S A *93*, 9061-6.

Goodall, G.J. and Filipowicz, W. (1989). The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell *58*, 473-83.

Goodall, G.J. and Filipowicz, W. (1991). Different effects of intron nucleotide composition and

secondary structure on pre-mRNA splicing in monocot and dicot plants. EMBO J *10*, 2635-44.

Guigo, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. (2000). An assessment of gene prediction accuracy in large DNA sequences. Genome Res *10*, 1631-42.

Guigo, R., et al (2003). Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. Proc Natl Acad Sci U S A *100*, 1140-5.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhar, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biology *3*, research 0029.1-0029.12

Hardison, R.C., Oeltjen, J., and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res *7*, 959-66.

Hastings, M.L. and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. Curr Opin Cell Biol *13*, 302-9.

Huang, X., Adams, M.D., Zhou, H., and Kerlavage, A.R. (1997). A tool for analyzing and annotating genomic sequences. Genomics *46*, 37-45.

Hubbard, T., et al (2002). The Ensembl genome database project. Nucleic Acids Res *30*, 38-41.

Karolchik, D., et al (2003). The UCSC Genome Browser Database. Nucleic Acids Res *31*, 51-4.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-64.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Ko, C.H., Brendel, V., Taylor, R.D., and Walbot, V. (1998). U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize. Plant Mol Biol *36*, 573-83.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. (2001). Integrating genomic homology into gene structure prediction. Bioinformatics *17 Suppl 1*, S140-8.

Lan, T.H., DelMonte, T.A., Reischmann, K.P., Hyman, J., Kowalski, S.P., McFerson, J., Kresovich, S., and Paterson, A.H. (2000). An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res *10*, 776-88.

Levine, A. and Durbin, R. (2001). A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res *29*, 4006-13.

Liu, H., Sachidanandam, R., and Stein, L. (2001). Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. Genome Res *11*, 2020-6.

Lykke-Andersen, J., Aagaard, C., Semionenkov, M., and Garrett, R.A. (1997). Archaeal introns: splicing, intercellular mobility and evolution. Trends Biochem Sci *22*, 326-31.

Makalowska, I., Ryan, J.F., and Baxevanis, A.D. (2001). GeneMachine: gene prediction and sequence

annotation. Bioinformatics *17*, 843-4.

Mathe, C., Sagot, M.F., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res *30*, 4103-17.

Miller, W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics *17*, 391-7.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. (1999). Frequent alternative splicing of human genes. Genome Res *9*, 1288-93.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet *30*, 13-9.

Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci *13*, 477-8.

Murakami, K. and Takagi, T. (1998). Gene recognition by combination of several gene-finding programs. Bioinformatics *14*, 665-75.

Mural, R.J., et al (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science *296*, 1661-71.

Novichkov, P.S., Gelfand, M.S., and Mironov, A.A. (2001). Gene recognition in eukaryotic DNA by comparison of genomic sequences. Bioinformatics *17*, 1011-8.

Ogasawara, J. and Morishita, S. Fast and sensitive algorithm for aligning ESTs to human genome. Proceedings of the First IEEE Computer Society Bioinformatics Conference. pp. 43-53. 2002.

Parra, G., Blanco, E., and Guigo, R. (2000). GeneID in Drosophila. Genome Res *10*, 511-5.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouzé, P. (1999). Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. Bioinformatics *15*, 887-99.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. (2000a). Genome annotation assessment in Drosophila melanogaster. Genome Res *10*, 483-501.

Reese, M.G., Kulp, D., Tammana, H., and Haussler, D. (2000b). Genie--gene finding in Drosophila melanogaster. Genome Res *10*, 529-38.

Sakata, K., et al (2002). RiceGAAS: an automated annotation system and database for rice genome sequence. Nucleic Acids Res *30*, 98-102.

Salamov, A.A. and Solovyev, V.V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res *10*, 516-22.

Salse, J., Piegu, B., Cooke, R., and Delseny, M. (2002). Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res *30*, 2316-28.

Tarn, W.Y. and Steitz, J.A. (1996a). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell *84*, 801-11.

Tarn, W.Y. and Steitz, J.A. (1996b). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science *273*, 1824-32.

Trachtulec, Z. and Forejt, J. (2001). Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. Mamm Genome *12*, 227-31.

Usuka, J. and Brendel, V. (2000). Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J Mol Biol *297*, 1075-85.

Usuka, J., Zhu, W., and Brendel, V. (2000). Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics *16*, 203-11.

Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. (2000). Human-mouse genome comparisons to locate regulatory sites. Nat Genet *26*, 225-8.

Wheelan, S.J., Church, D.M., and Ostell, J.M. (2001). Spidey: a tool for mRNA-to-genomic alignments. Genome Res *11*, 1952-7.

Xu, Y. and Uberbacher, E.C. (1996). Gene prediction by pattern recognition and homology search. Proc Int Conf Intell Syst Mol Biol *4*, 241-51.

Xuan, Z., Wang, J., and Zhang, M.Q. (2003). Computational comparison of two mouse draft genomes and the human golden path. Genome Biol *4*, R1

Yeh, R.F., Lim, L.P., and Burge, C.B. (2001). Computational inference of homologous gene structures in the human genome. Genome Res *11*, 803-16.

Zhu, W. and Brendel, V. (2002). Gene structure identification with MyGV using cDNA evidence and protein homologs to improve ab initio predictions. Bioinformatics *18*, 761-2.

Zhu, W., Schlueter, S.H., and Brendel, V. (2003). Refined annotation of the *Arabidopsis thanliana* genome by complete EST mapping . Plant Physiology *In press*.

CHAPTER 2. GENE STRUCTURE PREDICTION FROM CONSENSUS SPLICED ALIGNMENT OF MULTIPLE ESTS MATCHING THE SAME GENOMIC LOCUS

A paper to be submitted to *Proceedings of the National Academy of Sciences*

Wei Zhu[1] and Volker Brendel[2]

Abstract

Accurate gene structure annotation is a challenging computational problem in genomics. Best results are achieved with spliced alignment of full-length cDNAs or multiple ESTs with sufficient overlap to cover the entire gene. For most species, cDNA and EST collections are far from comprehensive. We have developed a computer program, GeneSeqer, which is capable of aligning thousands of ESTs with a long genomic sequence in a reasonable amount of time (available at http://bioinformatics.iastate.edu/cgi-bin/gs.cgi). The algorithm is uniquely designed to tolerate a high percentage of mismatches and insertions or deletions in the EST relative to the genomic template. This feature allows use of non-cognate ESTs for gene structure prediction, including ESTs derived from duplicated genes and homologous genes from related species. We assessed GeneSeqer performance relative to a standard *Arabidopsis thaliana* gene set and demonstrate its utility for plant genome annotation. In particular, we propose that this method provides a much needed tool for the annotation of the rice genome, using abundant ESTs from other cereals and plants.

Introduction

Annotation of gene structure in eukaryotic genomes currently involves both computational and experimental approaches. Because of time and expense constraints, initial annotation mostly relies on *ab initio* gene prediction based on statistical modeling of exon and intron features. The best of these methods have been estimated to achieve about 80% sensitivity and specificity at the exon level, but

---

[1] Primary researcher and author, graduate student, Department of Zoology and Genetics, Iowa State University.

[2] Author for correspondence, Professor, Department of Zoology and Genetics, Department of Statistics, Iowa State University

the success rate is much lower at the level of entire gene structure, with typically less than half the predictions entirely accurate (1, 2). In practice, a combination of different programs appears to be more successful than reliance on a single program (1, 3). Spliced alignment of potential homologous protein sequences to genomic DNA is a complementary approach to *ab initio* gene prediction that gives better accuracy, provided a close enough homolog of the potential gene product is available (4-6).

The most direct experimental evidence for gene structure comes from sequencing full-length cDNAs with subsequent spliced alignment of the cDNA sequences to the genomic DNA. An added advantage of this approach is that sufficient cDNA sampling under different conditions will reveal transcript isoforms arising from alternative splicing or alternative transcription start or termination points. An intermediate step in gene discovery is sequencing of expressed sequence tags (ESTs), which typically correspond to partial rather than full-length cDNAs. Clustering and assembly of ESTs to potential full-length transcripts is commonly pursued to estimate the gene space of a species, using methods that rely on pair-wise sequence similarities (7-10). However, direct alignment to genomic DNA, when possible, is more accurate and informative (11).

The alignment of ESTs to genomic DNA is non-trivial for a number of reasons. ESTs are usually deposited as single-pass sequencing products, increasing the conventionally accepted rate of sequencing errors and ambiguous base determinations. ESTs are typically sampled from a large variety of origins that represent a range of subspecies, tissue types, and conditions, thus leading to a heterogeneous sequence view confounded by polymorphisms and paralogous genes. In addition, sequencing artifacts (e.g., chimeras), sample contaminations, and complex patterns of alternative splicing further complicate the alignment task.

A number of tools that address this alignment problem are now available and provide adequate solutions for some of these needs in more narrowly defined context. The underlying algorithms can be categorized into two groups with respect to the way they generate spliced alignments. One category involves heuristic, BLAST-like methods for the initial alignment and includes the tools sim4 (12), Spidey (13), BLAT (14), and Squall (15). Typically, these programs find matching segments at high stringency using BLAST (16) or a variant, with subsequent output parsing to favor canonical splice sites. EST_GENOME (17), dds/gap2 (18) and GeneSeqer (19) belong to another category of programs that implement a full dynamic programming approach to derive the optimal score and spliced alignment, allowing for within-exon insertions and deletions. In GeneSeqer, potential splice sites are differentially scored according to independent splice site prediction methods. Consideration of predicted splice site strength was shown to improve the performance of the algorithm in the case of

imperfect sequence matching as a result of sequencing errors or alignment of non-cognate, but homologous ESTs (19).

There are several limitations in the BLAST-like spliced alignment methods. First, short exons (about 20 or fewer bases) are generally missed because they do not qualify as high-scoring segment pairs. Second, reliable alignments are limited to cognate ESTs with low sequencing error rates. For example, sim4 reports only the highest scoring match for each EST query, and TAP, a useful transcript assembly tool based on sim4 (20), recommends a threshold of 92% overall identity for any such alignment to be included into the transcript assembly. In addition, the simple adjustment for exon-intron boundaries to conform to canonical splice sites whenever possible, as used in most of spliced alignment programs, further restricts application to unequivocal alignments and can lead to inconsistencies (e.g., sim4/TAP allow the standard GT-AG introns in conjunction with a complementary CT-AC intron in the same alignment, confounding assignment of the true transcript orientation). These limitations may be inconsequential when the need is for fast, reliable alignment of ESTs or cDNAs that, based on high sequence similarity, can be unambiguously assigned to a unique chromosomal locus, however they render these algorithms helpless in the situations discussed here.

EST sampling is sparse for most species when compared with the large human and mouse EST collections. However, if ESTs from related taxonomic groups could be successfully employed for gene identification, the EST resources would appear much more impressive. To date, there are well over two million ESTs from all plant species combined. Because of the inclusion of sophisticated splice site models and exhaustive alignment with a dynamic programming approach, the GeneSeqer algorithm affords a promising approach in attempts to make use of this resource. For example, GeneSeqer was recently shown to be very successful in identifying very short exons in *Arabidopsis thaliana* (21) and improving *Arabidopsis* genome annotation (11). Here we report generalization of GeneSeqer to exploit heterogeneous EST sources for plant genome annotation.

The greater accuracy afforded by the dynamic programming approach adopted in GeneSeqer is obtained at the expense of greater computational efforts. Practical implementation of the algorithm requires efficient selection of restricted genomic DNA regions and matching ESTs from a typically large EST collection in order to minimize or eliminate the computer time spent on deriving locally optimal, but insignificant alignments. In this study, we present a string matching scheme based on pre-processing of the input EST data set that allows fast target selection for detailed analysis by the dynamic programming algorithm. The current GeneSeqer algorithm was also modified to incorporate results of Bayesian statistical models for splice site prediction described elsewhere (22). We discuss

applications to *Arabidopsis thaliana* and rice genome annotation, which suggest that this approach provides a practical and powerful tool for accurate gene structure identification.

Materials and Methods

Programs used. The dynamic programming subroutines of GeneSeqer were described previously (5, 19). The source code of the program is available at http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. The data and some of the figures in this article were produced with the specialized GeneSeqer Web servers at http://www.plantgdb.org/cgi-bin/AtGeneSeqer.cgi (for *Arabidopsis*) and http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi (all plant species; ref. 23). Sim4 (12) was downloaded from http://globin.cse.psu.edu/. TAP (20) was obtained from http://sapiens.wustl.edu/~zkan/TAP/. The Spidey (13) executable was obtained from http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/spideyexec.html. The BLAT (14) executable was compiled from the source code made available at Jim Kent's Web page, http://www.soe.ucsc.edu/~kent/src/.

Spliced threading. The GeneSeqer algorithm solves the problem of "threading" an EST or cDNA into a genomic DNA sequence such that each nucleotide in the matching genomic DNA segment is consistently assigned exon or intron status (for example, Fig. 1). The threading preferentially selects high-scoring splice sites unless strongly contradicted by sequence similarity supporting lower scoring sites. An optimal alignment score is calculated by dynamic programming as described previously (19). In similar fashion, GeneSeqer also derives the optimal threading of a protein sequence onto the inferred translation of a genomic DNA segment, allowing gene prediction by similarity to putative homologs of the given locus (5).

Scoring. A number of parameters influence the optimal alignments, including standard scores for identities, mismatches, and deletions within exon alignments. In addition, persistence within and switching between exon and intron states is governed by transitions probabilities derived from splice site prediction values along the genomic sequence (5, 19). These values are calculated for all positions in the genomic sequence prior to the spliced alignment. Precisely, default donor site values are assigned as 0.00005 for any GT and 0.00002 for any GC or AT (similarly, 0.00005 for any AG and 0.00002 for any AC as potential acceptor sites). The other dinucleotides have a default score 0.000001 as donor or acceptor site value. These default values are replaced by $2\times(P-0.5)$, where P is

the respective Bayesian a posteriori splice site probability, whenever that value is greater. Empirically, this scaling seems to give a good balance between scoring for sequence similarity and scoring for splice site consensus (the balance can easily be changed by providing the GeneSeqer program at run time with other than default parameters).

The quality of a particular optimal alignment is assessed by similarity and coverage scores. The similarity score is calculated as a normalized alignment score and is derived for each exon, the 50-base exon flanks of each predicted intron, and for the entire alignment by averaging over all exons of at least 50 bases (e.g., Fig. 2). Note, that with default parameters, in the absence of insertions/deletions a similarity score of s would correspond to $0.5\times(1+s)\times100\%$ sequence identity. The coverage score gives the length of the matching region relative to the entire EST length (i.e., a completely matched EST would have coverage score 1.0).

Quality adjustments. By default, GeneSeqer will align any EST to a genomic locus with which it shares at least partial significant similarity as determined in the fast screen for matching loci described below. This may result in optimally scoring, but clearly poor alignments over the entire EST when the significant similarity is limited to disjoint segments of the EST. While such alignments can still be useful to indicate exon potential in the matching genomic segments (if not an entire gene structure), the GeneSeqer program also provides a post-processing step that quality-adjusts such alignments based on user-specified parameters. Briefly, a predicted gene structure is assessed exon by exon, starting with the terminal exons, with weakly matching terminal exons recursively being eliminated. The elimination process involves a decision tree. For example, the 3′-most exon in a multiple-exon predicted gene structure is quality-adjusted as follows: 1) Is the exon score below the parameter POOR_EXON_SCORE (default: 0.7)? If yes, and 2) the exon length is at most TINY_EXON bases (default: 20), the exon is removed from the alignment. If the exon is longer, and 3) the acceptor site score is at most POOR_ACPTR_SCORE (default: 0.5) or 4) the length of the intron is at least LONG_INTRON bases (default: 300), the exon is removed. If conditions 3) and 4) for elimination are not met, the exon is retained unless the upstream exon is to be eliminated by the same criteria. To complete the decision tree, exons that 1) score above POOR_EXON_SCORE are retained if, 2), they are of length greater than TINY_EXON. However, they are eliminated if they are shorter and successively either 3) the acceptor score is poor, 4) the intron is long, or the upstream exon is weak. Predicted 5′ gene structure ends are similarly adjusted.

Strand selection. Based on sequence similarity alone, a spliced alignment could be made equally with either strand of a genomic DNA. For multi-exon alignments, GeneSeqer orients the alignment to maximize the average splice site score. For example, the alignment in Fig. 2A assigns high splice site scores for introns two and three, thus aiding in the detection of the first intron as an AT-AC intron in the same orientation. In ambiguous cases, occasional retention of a poly-A tag in the EST sequence may indicate the direction of transcription. For single exon alignments, GeneSeqer assigns a putative transcription orientation based on overlap with multi-exon alignments as described next. In general, no attempt is made to use annotated orientation, if available, because we have found such annotation not always reliable. However, a particular alignment orientation can be enforced at run time.

Consensus gene structures. A critical step in our strategy to predict gene structure by spliced alignment is the derivation of a consensus gene structure prediction from multiple, possibly low scoring, overlapping spliced alignments. If the resulting gene structure spans multiple exons and contains an open reading frame across these multiple exons, confidence in the prediction should be very high, because the GeneSeqer algorithm (unlike *ab initio* gene prediction programs) does not score in any way for coding frame consistency. Fig. 1 provides a typical example, discussed below.

The determination of consensus gene structures in our algorithm is a multi-step process. First, all EST alignments are clustered into Predicted Gene Locations (PGLs) based on genomic location. This clustering is achieved by going through all the alignments by increasing left-point coordinate. Clusters are separated by gaps of at least JOIN_LENGTH bases, a parameter that can be changed at run time (default: 300). An exception to this is made if a new alignment is of opposite orientation compared to the current PGL; in this case, a new PGL is assigned. Single exon alignments are displayed in the orientation of their associated PGL. If a PGL consists entirely of single exon alignments, then the orientation is determined first by the presence of any potential poly-A tags and second by choosing the orientation that gives the longest open reading frame. It is clear that intergenic regions less than JOIN_LENGTH may cause problems, but empirically these rules seem to work very well (see ref. 11 for extensive applications to *Arabidopsis*).

Within each PGL, alternative splicing would result in inconsistent predicted gene structures (PGSs) from individual ESTs. This is represented in the GeneSeqer output by multiple alternative gene structures (AGSs) within a single PGL. An example is given in Fig. 3. Assembly of AGSs proceeds left to right, with each PGS added into the current AGS as long as its exon/intron assignments are consistent with the current AGS. Otherwise, a new AGS is started. The alignment ends of an AGS may be slightly adjusted to fit a PGS. This adjustment eliminates wrong alternative

splicing predictions that would otherwise result from weak, random matching of EST end sequences, which are typically of lower sequence quality. The GeneSeqer output only indicates the alternative transcript isoform fragments confirmed by spliced alignment but does not further process these fragments to assemble all potential full-length transcript isoforms. However, the output could easily be parsed and re-formatted for input into the TAP program (20) for this purpose (currently, TAP uses sim4 spliced alignments by default).

Fast screen for matching ESTs. Efficient use of EST evidence for genome annotation requires mapping large EST collections onto BAC-size genomic DNA segments. Because dynamic programming is computationally prohibitive for such large problems, a fast screen must be implemented to select promising EST matches for gene-sized genomic segments. In the absence of very long introns, the dynamic programming algorithm can then be applied to the selected DNA input (the case of long introns can be handled by more sophisticated screening that eliminates presumed intron-internal sequences; not pursued here). For GeneSeqer, we have implemented the suffix array method of Manber & Myers (24) for pre-processing of the EST database. Note that for applications in which the genomic DNA query is fixed (e.g., annotation of a complete genome), additional pre-processing of the genomic DNA sequence may be considered.

Two parameters determine the outcome of the initial screen for matching ESTs. The GeneSeqer -*x wsize* option specifies the minimal exact match size for successful extension (typically, *wsize* is set to 12 to 16; higher values allow much faster screening for high quality matches only). Precisely, the genomic DNA query is processed from in 5' to 3' direction, with each consecutive *wsize*-mer match against the EST database added to a set of linked lists that store match information for each specific EST. As the linked lists grow, the matches from each individual EST are continuously merged into high-scoring segment pairs (HSPs) that allow for small insertions and deletions in both genomic DNA and EST. Related HSPs are then further chained together to define matching regions between the genomic DNA and the specific EST using the algorithm of Pearson and Lipman (25), with minor penalty for long gaps in the genomic region (possible introns). These two steps are analogous to the first two steps in the algorithm applied by sim4 (12). However, sim4 only utilizes the best scoring chain for each EST, whereas multiple non-intersecting chains with significant scores higher than a cutoff value would be selected in GeneSeqer. This allows a single EST to be matched to different genomic loci. This property is crucial for the applications discussed here. The cognate EST location is easily identified as the highest scoring match, but, in addition, an EST can often be successfully used to identify gene structure in a duplicated locus, in particular a locus with potentially low cognate

EST representation (11). The cutoff value for successful HSP chains is specified by the GeneSeqer *-y minQ* argument. Each promising region is then slightly expanded to allow for uncertainties at the ends, and the full dynamic programming alignment is applied to this genomic DNA region and the entire EST sequence.

Complexity. A typical application of GeneSeqer is to map a large EST collection (total sequence length M) to a single genomic sequence of length n. The whole process of EST mapping consists of three parts: construction of the suffix array for the EST sequences, genomic localization (fast screen with GeneSeqer option *–x wsize*), and spliced alignment. The run time for building the suffix array for ESTs is $O(M \cdot log\ M)$ (24). This computational time is typically negligible because a large number of ESTs are usually preprocessed to build the suffix array, which avoids potential overhead in repeated small-scale analyses. The genomic localization step is very fast with run time $O(n \cdot (wsize + log\ M))$, based on a search algorithm for suffix arrays using longest common prefixes (26). Therefore, the computation for large-scale mapping is dominated by the cost for the spliced alignment part and thus is linearly proportional to the expected number of alignments and the square of the average alignment length.

Evaluation. To benchmark the prospects and limits of gene prediction by spliced alignment, we evaluated the GeneSeqer performance on the AraSet *Arabidopsis* gene set distributed for such purposes by Pavy et al. (1), available at http://sphinx.rug.ac.be:8080/biocomp/napav/. This set consists of 74 contigs comprising two to four genes each, 168 genes and 859 introns in total. Spliced alignments were based on the mapping of 176,195 *Arabidopsis* ESTs that were downloaded from the NCBI dbEST database (27).

To evaluate prediction accuracy at the intron level, we define correct introns, overlapping introns, wrong introns and missed introns as in (1). Thus, a predicted intron identical to an annotated intron is classified as "correct intron". An "overlapping intron" refers to a predicted intron overlapping with some annotated intron, but with different 5' and/or 3' splice site. A "wrong intron" refers to a predicted intron overlapping with annotated exons, but not with annotated introns. Both overlapping introns and wrong introns are counted as incorrect (false positive) predictions (note that this assumes lack of alternative splicing in the test set). "Missed introns" are annotated introns that are not overlapped by any predicted intron (false negatives). Because only introns in coding sequences (CDS) are annotated in AraSet, introns predicted by spliced alignment outside of CDS cannot be evaluated. Thus, sensitivity at the intron level is defined as (number of correct introns / number of

annotated introns), and specificity is determined as (number of correct introns / number of predicted introns in CDS).

Results and Discussion

Spliced alignment with heterologous ESTs. Fig. 1 illustrates the application of spliced alignment for gene structure annotation. The upper panel shows nine PGSs with rice ESTs (red) that result in three disjoint AGSs (green). The complete alignments are available as Supporting Information at http://www.plantgdb.org/AtGDB/prj/ZB03PNAS/atac/gs_sorted-output-1049503986_7781_top.html. The three AGSs are supported by similarity scores of about 0.8, 0.9, and 0.95, respectively. While the 3′-terminal exons of the annotated gene structure are confirmed by spliced alignment, contradictory results are obtained at the 5′-end. This issue is resolved when ESTs from plants other than rice are added. Using the GeneSeqer Web service at PlantGDB (23), a total of 266 ESTs could be significantly aligned in this region. The lower panel in Fig. 1 depicts the results for a subset of these ESTs, all derived from barley. Several of these ESTs bridge all coverage gaps and predict a single gene structure with seven exons (green). An open reading frame (orange) spans all the exons, and its translation identifies the gene as coding for a UDP-glucuronic acid decarboxylase. Of particular interest is identification of the first intron (785 bases) as a U12-type intron with AT-AC borders (Fig. 2A). The intron is in a coding region that is highly conserved with an *Arabidopsis* homolog (Fig. 2B), and the *Arabidopsis* gene also has a U12-type intron in the same position (although none but the U12 signatures are preserved in the intron sequences; see ref. 11).

As shown in Fig. 2A, by including splice site scoring and preferences, GeneSeqer can use even quite diverged ESTs to predict the correct gene structure (in this case, the barley EST have an average similarity score of only 0.72). For comparison, none of the other programs we tried (sim4, blat, Spidey) produced any alignments for the same genomic DNA and EST.

Evaluation of spliced alignment accuracy. We have recently reported on the utility of spliced alignment to correct and refine *Arabidopsis thaliana* genome annotation (11). As an independent assessment of the applicability and performance quality of GeneSeqer, here we evaluate its accuracy relative to the AraSet test set compiled by Pavy et al. (1). All available *Arabidopsis* ESTs were mapped onto the AraSet contigs using GeneSeqer default parameters. Post-screening of the reported alignments was used to select subsets of alignments satisfying more stringent match criteria. Because the alignments with terminal ESTs correspond to predicted transcript ends rather than coding region ends as in the AraSet annotation, evaluations were made entirely on the intron level, using standard

performance measures (1).

Results are summarized in Table 1. With default parameters, the spliced alignment indicated 782 introns (compared to 859 annotated introns in AraSet). Of these, 625 introns coincided with annotated introns for a sensitivity of 0.728. Assessment of specificity is less straightforward. First, spliced alignment, unlike *ab initio* programs tested on AraSet, can reveal introns in untranslated regions (UTRs). Here, 76 introns were predicted outside of the CDS bounds annotated in AraSet. Careful inspection indicated that this set contains both UTR introns and introns of genes that were omitted in the AraSet annotation (see below). A second problem is that some of the overlapping introns may correspond to correctly predicted alternative transcripts. Thus, the listed specificity of 0.885 may be underestimating the actual specificity.

In order to clearly separate errors of the spliced alignments from errors in the AraSet annotation, we evaluated a subset of all predicted introns that satisfy very stringent alignment quality criteria. Let *Pd* (*Pa*) and *Sd* (*Sa*) denote the splice site score and local similarity score for each donor (acceptor) site, respectively (*cf.* Fig. 2). Requiring $Pd>0$, $Pa>0$ and $Sd>0.95$, $Sa>0.95$ selects only introns with canonical splice sites supported by EST matching with more than 97.5% identity in the flanking 50 exon bases. For this subset, 463 of the 471 predicted introns within CDS bounds coincide with the AraSet annotation. The remaining eight introns were further scrutinized, and all seem authentic (Table 2). In three cases (seq53, seq62 and seq72), the annotated introns are supported by other ESTs, thus the two conflicting coordinate sets represent alternative splicing events. In the other five cases, there is no EST support for the annotation, and thus the EST-supported coordinates may be assumed to be the correct annotation. With that correction, the specificity of GeneSeqer high quality intron prediction is 100%, as expected. Sensitivity in this case dropped to just over 50%. For comparison, exon level sensitivity and specificity were estimated at just above and below 80%, respectively, for the best *ab initio* gene prediction programs (1).

Sensitivity for the spliced alignment approach depends mostly on the availability of ESTs. However, when using non-cognate ESTs, we are also assessing the ability of the program to use such data for accurate prediction. As displayed in Table 1, with GeneSeqer default parameters a gain of about 20% in sensitivity is accompanied by a drop in specificity of about 10%. Restriction of the predicted intron set to only canonical introns (without the additional requirement for high quality flanking exon matching) gives intermediate values.

There were 28 introns in the high quality subset that are not located within the annotated CDS bounds, and thus are potential UTR introns. Further analysis indicates that some of these introns are actually from genes that are not annotated in AraSet. For example, three genes are annotated in

AraSet contig seq25, with a 4.4 kb "intergenic region" between the second and the third gene. The most recent *Arabidopsis* genome annotation suggests that there is a gene At5g63670 with five exons in the "intergenic region", supported by three full-length cDNAs and three ESTs. Similar situations also occur in the AraSet contigs seq30, seq41, and seq69. Supporting data for all these cases are available as Supporting Information at http://www.plantgdb.org/AtGDB/prj/ZB03PNAS/AraSet/AraSet-AtGDB.html.

Applications to rice genome annotation. To test the utility of GeneSeqer for the annotation of the rice genome, we analyzed a randomly selected rice BAC (GenBank accession AP002487) in detail. Spliced alignment results of the central 44,000 bases of the sequence are displayed as Supporting Information at http://www.plantgdb.org/AtGDB/prj/ZB03PNAS/OsBAC/gs_sorted-output-1049500973_7289_top.html. Overall, spliced alignment confirmed six genes, only one of which agrees with the current gene annotation provided in the GenBank file. Fig. 4 summarizes the results for two adjacent genes. In both cases, a sufficient number of ESTs from heterogeneous sources could be found to give a complete tiling of the gene, supported by open reading frames spanning all exons and showing high similarity to known *Arabidopsis* gene products.

Conclusions. After genome sequencing and assembly, genome annotation is the critical task in the characterization of the genetic blueprint of an organism. For all eukaryotic model organisms that have been sequenced, the annotation efforts have continued and are continuing for years after the initial sequence release. Thus, the human genome is still being evaluated, and in particular, the abundance of alternative splicing of human genes has only recently been appreciated (28-30). The annotation tasks for plant genomes currently pose distinct challenges compared to vertebrate genome annotation. First, EST and full-length cDNA availability is much less for plants than for human and mouse. Currently, there are 416,000 wheat ESTs as the largest plant collection, compared to more than five million for human and 3.7 million for mouse (see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Only 131,000 rice ESTs are publicly available, less than the 179,000 *Arabidopsis* ESTs for an about threefold smaller genome. Secondly, all plant genomes surveyed to date are replete with gene duplications as a result of both polyploidization and random segmental duplications (e.g., see ref. 31, 32).

We have recently reported the mapping of all *Arabidopsis* ESTs onto the *Arabidopsis* genome using GeneSeqer and showed that about 65% of annotated gene locations had EST evidence, with full coverage for about 23% of the genes (11). Here we have presented details of the GeneSeqer

algorithm with respect to the derivation of consensus gene structures from multiple ESTs from potentially hetereogeneous, diverged sources. A number of key differences in the algorithm compared to other programs geared towards fast alignment of cognate ESTs allow efficient use of non-native EST resources. We believe this will greatly aid in the annotation of plant genomes, particularly rice and maize. The GeneSeqer Web service at PlantGDB (http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi) should allow any member of the plant research community easy access to the annotation tools, and we hope that such community input will quickly improve the status of plant genome annotation.

Acknowledgments

We would like to thank Dr. Qunfeng Dong and Shannon D. Schlueter for critical reading of the manuscript and helpful discussions. V.B. was supported in part by NSF grants DBI-9872657 and DBI-0110254.

References

1. Pavy, N., Rombauts, S., Dehais, P., Mathé, C., Ramana, D. V., Leroy, P. & Rouzé, P. (1999) *Bioinformatics* **15**, 887-899.

2. Rogic, S., Mackworth, A. K. & Ouellette, F. B. (2001) *Genome Res.* **11**, 817-832.

3. Murakami, K. & Takagi, T. (1998) *Bioinformatics* **14**, 665-675.

4. Gelfand, M. S., Mironov, A. A. & Pevzner, P. A. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 9061-9066.

5. Usuka, J. & Brendel, V. (2000) *J. Mol. Biol.* **297**, 1075-1085.

6. Mathé, C., Sagot, M. F., Schiex, T. & Rouzé, P. (2002) *Nucleic Acids Res.* **30**, 4103-4017.

7. Bouck, J., Yu, W., Gibbs, R. & Worley, K. (1999) *Trends Genet.* **15**, 159-162.

8. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. & Quackenbush, J. (2000) *Nucleic Acid Res.* **28**, 3657-3665.

9. Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. & Quackenbush, J. (2003) *Bioinformatics* **19**, 651-652.

10. Kalyanaraman, A., Aluru, S., Kothari, S., & Brendel, V. (2003) Efficient clustering of large EST

data sets on parallel computers. *Nucleic Acids Res.*, in press.

11. Zhu, W., Schlueter, S. D. & Brendel, V. (2003) Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiology*, to appear June issue.

12. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. (1998) *Genome Res.* **8**, 967-974.

13. Wheelan, S. J., Church, D. M. & Ostell, J. M. (2001) *Genome Res.* **11**, 1952-1957.

14. Kent, W. J. (2002) *Genome Res.* **12**, 656-664.

15. Ogasawara, J. and Morishita, S. (2002) in *Proceedings of the First IEEE Computer Society Bioinformatics Conference*. (Stanford, California), pp. 43-53.

16. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389-3402.

17. Mott, R. (1997) *Comput. Appl. Biosci.* **13**, 477-478.

18. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. (1997) *Genomics* **46**, 37-45.

19. Usuka, J., Zhu, W. & Brendel, V. (2000) *Bioinformatics* **16**, 203-211.

20. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. (2001) *Genome Res.* **11**, 889-900.

21. Haas, B. J., Volfovsky, N., Town, C. D., Troukhar, M., Alexandrov, N., Feldmann, K. A., Flavell, R. B., White, O. & Salzberg, S. L. (2002) *Genome Biol.* **3**, research 0029.1-0029.12.

22. Xing, L. & Brendel, V. (2003) Species-specific splice site recognition by sequence inspection using Bayesian statistical models. *Nucl. Acids Res.*, submitted.

23. Schlueter, S.D., Dong, Q. & Brendel, V. (2003) GeneSeqer@PlantGDB - gene structure prediction in plant genomes. *Nucl. Acids Res.*, in press.

24. Manber, U. & Myers, G. (1993) *SIAM J. Comput.* **22**, 935-948.

25. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448.

26. Gusfield, D. (1997) in *Algorithms on strings, trees, and sequences: Computer Science and Computational Biology*. (Cambridge University Press, New York), pp. 152-155.

27. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. (1993) *Nat. Genet.* **4**, 332-333. http://www.ncbi.nlm.nih.gov/dbEST/.

28. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. (1999) *Genome Res.* **9**, 1288-1293.

29. Modrek, B. & Lee, C. (2002) *Nat. Genet.* **30**, 13-19.

30. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. (2002) *Nat. Genet.* **30**, 29-30.

31. Gaut, B. S. (2001) *Genome Res.* **11**, 55-66.

32. Blanc, G., Hokamp, K. & Wolfe, K. H. (2003) *Genome Res.* **13**, 137-144.

Figure Legends

Figure 1. Gene structure annotation for a putative rice gene on chromosome one. The schematic displays of the GeneSeqer spliced alignments were generated with the GeneSeqer Web server at the PlantGDB site (http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi; ref. 23). The scale refers to the numbering of the BAC sequence deposited in GenBank as accession AP003271. GenBank CDS annotation is shown in light blue, with solid boxes corresponding to exons and thin lines corresponding to introns. The arrow indicates the direction of transcription. The same convention is used for EST spliced alignments (red), alternative gene structures (green) derived from consistently overlapping EST spliced alignments, long open reading frames (orange), and protein spliced alignments (purple). **Upper Panel:** Spliced alignment of nine rice ESTs confirms the five 3'-most exons of the annotated gene structure, but is inconclusive with respect to the 5'-end of the gene. **Lower Panel:** Spliced alignment with 53 barley ESTs suggests a seven-exon gene structure (green), which encodes a single long open reading frame (orange). The translation product is highly similar to the *Arabidopsis* gene At3g53520 product, a UDP-glucuronic acid decarboxylase, and direct spliced alignment of the *Arabidopsis* protein supports the same gene structure (purple; see also Fig. 2B). A protein database search showed that the rice homolog has been deposited as GenBank accession BAB84333.

Figure 2. **A.** Partial GeneSeqer output of the spliced alignment of barley EST gi:21142201 (plus strand) with the rice locus represented by BAC AP003271, 154,000 to 158,000 region. The spliced alignment predicts four exons as summarized on top. Exon scores are normalized sequence similarity scores. Pd, donor site score; Pa, acceptor site score (the s-values in parenthesis are the normalized sequence similarity scores in the adjacent 50 exon nucleotides). The MATCH line shows the average similarity score, the matching sequence length, and the coverage score relative to the EST length (see METHODS for details). Only part of the alignment is shown (omitted parts indicated by //). Identities between the genomic sequence (upper lines) and the EST sequence (lower lines) are indicated by vertical bars. Introns are represented by dots. GeneSeqer correctly identified the 785-

base AT-AC intron, even though the EST alignment contains many mismatches (mostly in third codon positions). **B.** Spliced alignment results for the same rice locus with a homologous *Arabidopsis* protein (At3g53520). The alignment summary is as for the EST alignment. The sequence alignment shows high conservation on the protein level in the exons flanking the AT-AC intron.

Figure 3. Alternative gene structure prediction for AraSet entry seq62 representing the *Arabidopsis* At4g37070 gene. Symbols and colors are as explained in the legend to Figure 1. In addition, GenBank mRNA annotation is shown in dark blue. The eight matching ESTs (red) were assembled into three consistent transcript fragments (green). The first intron has two alternative donor sites, supported by two and three ESTs, respectively. Note that the GeneSeqer program does not attempt to display all possible full-length transcript isoforms. However, inspection of the open reading frames (orange) suggests that the gene may have two transcript isoforms differing only in the first donor site, but maintaining the reading frame such that the two protein isoforms differ only by an additional 11 amino acids in the longer protein.

Figure 4. Example of rice genome annotation by spliced alignment. Symbols and colors are as explained in the legend to Figure 1. The scale refers to the numbering of the BAC sequence deposited in GenBank as accession AP002487. The spliced alignments were generated with the GeneSeqer Web server at the PlantGDB site by aligning a total of 71 matching ESTs from all plant species (details available as Supplementary Information at http://www.plantgdb.org/AtGDB/prj/ZB03FNAS/OsBAC/gs_sorted-output-1049500973_7289_top.html. This figure shows the alignments of 30 representative ESTs only (red). The current GenBank annotation (light blue) for both genes is incorrect. The upstream gene encodes a homolog of the *Arabidopsis* gene At1g66680 product, a putative pheromone receptor, and the downstream gene encodes a homolog of the *Arabidopsis* gene At2g01275 product, a putative protein with similarity to nucleoside triphosphatase (protein spliced alignments shown in purple).

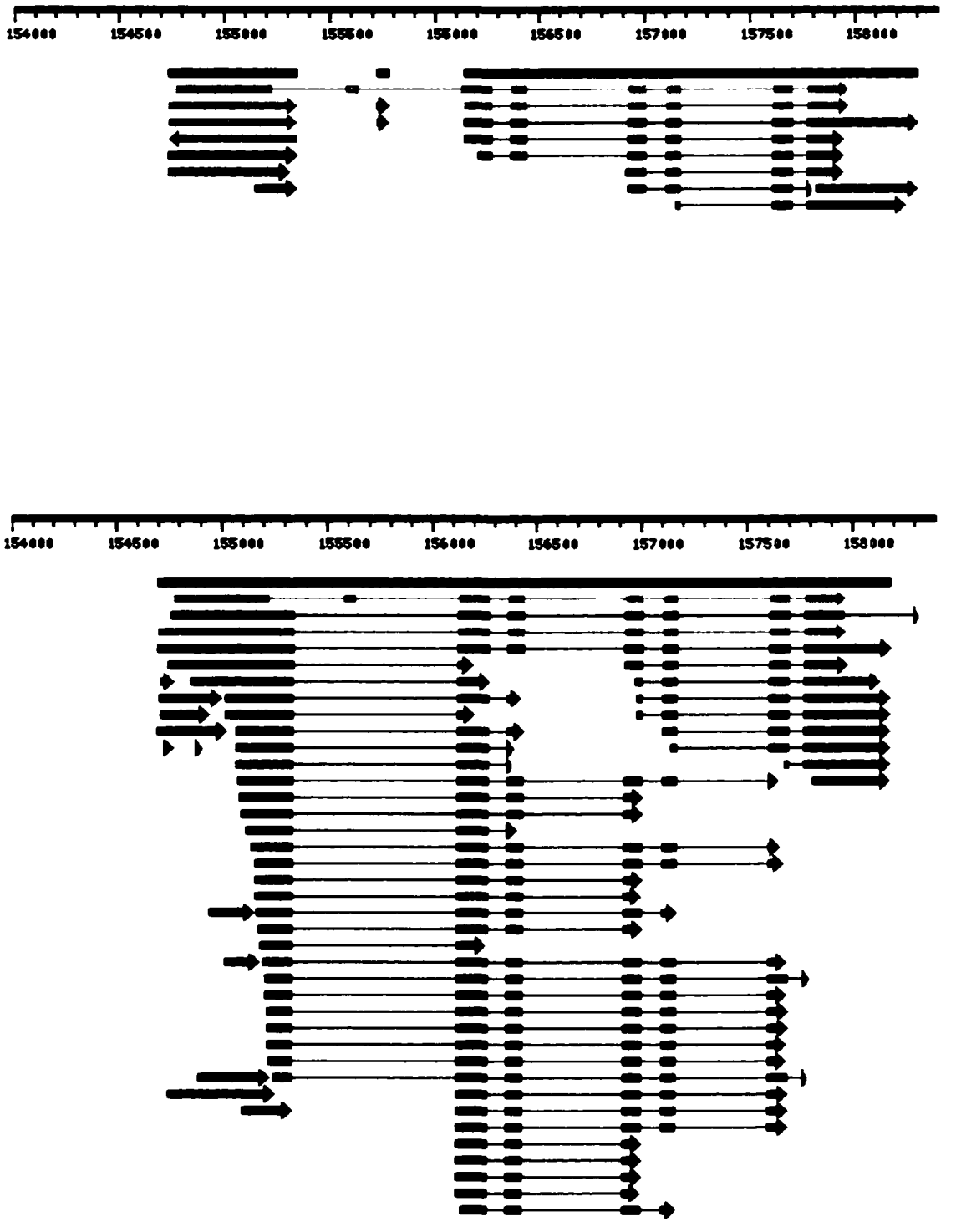

154000
154500
155000
155500
156000
156500
157000
157500
158000
154000
154500
155000
155500
156000
156500
157000
157500
158000

Figure 1

```
EST sequence gi:21142201+

Predicted gene structure (within gDNA segment 154657 to 157529):

 Exon  1 155165 155338 ( 174 n);  cDNA     84    257 ( 174 n); score: 0.701
  Intron  1 155339 156123 ( 785 n); Pd: 0.000 (s: 0.74), Pa: 0.000 (s: 0.80)
 Exon  2 156124 156271 ( 148 n);  cDNA    258    405 ( 148 n); score: 0.770
  Intron  2 156272 156360 (  89 n); Pd: 0.989 (s: 0.80), Pa: 0.987 (s: 0.66)
 Exon  3 156361 156440 (  80 n);  cDNA    406    485 (  80 n); score: 0.637
  Intron  3 156441 156922 ( 482 n); Pd: 0.999 (s: 0.58), Pa: 0.791 (s: 0.68)
 Exon  4 156923 157006 (  84 n);  cDNA    486    569 (  84 n); score: 0.738

MATCH   AP003271+      21142201+      0.718   486     0.817   C
PGS_AP003271+_21142201+       (155165  155338,156124  156271,156361  156440,
  156923  157006)

Alignment (genomic DNA sequence = upper lines):

//

GACAACTTCT TCACCGGGAG GAAGGACAAC GTCGCGCACC ACCTCCGGAA CCCCAGGTTC   155224
||||||||  |  ||||  ||      ||  ||||||   |    | |       ||  |   |  |||  ||  ||
GACAACTTTT TCACTGGTTC AAAAGACAAC CTGAAGAAGT GGATCGGCCA CCCAAGATTT      143


GAGCTGCTCC GCCACGATGT CGTCGAGCCC ATCCTGCTCG AGGTGGACCG GATCTATCAC   155284
|||||   |||  |  ||||||||  |   |||  ||    |  ||   |  |  ||||  ||||    |||||||||
GAGCTCATCC GTCACGATGT CACCGAACCA CTGCTCGTGG AGGTTGACCA AATCTATCAC      203


CTCGCGTGCC CCGCGTCCCC TGTGCACTAC AAGTACAACC CCATCAAGAC GATCATATCC   155344
||  ||  ||||  |  ||  ||  ||     |    ||||  |||  ||||||  |   |||||||   |||
CTTGCTTGCC CTGCCTCACC AATCTTCTAC AAGCACAACC CTGTCAAGAC CATC......      257

//

AATATAATCA TGTGTTGCAT CTGTATATTA TCAAATCTGG GTTTCTTAAC AATTATTACA   156124
                                                                   |
.......... .......... .......... .......... .......... .........A      258


AGACCAATGT CATGGGAACC TTGAATATGT TGGGTCTGGC AAAGCGAATT GGTGCAAGGT   156184
||||  |||||  |||  ||  |||  |||||  |||   |  ||  ||  ||  ||||  ||  ||  ||  ||  ||
AGACGAATGT CATTGGGACC TTGAACATGC TTGGACTTGC AAAGAGAGTT GGAGCTAGAA      318

//
```

Figure 2A

```
Query protein sequence: At3g53520.1

     1  MKQLHKQMSS KRDEETIPMS QSSPYSPKTL KHPRSLPRSL HYLFREQRLL FILVGILIGS
    61  TFFILQPSLS RLGAAESTSL ITRSVSYAVT DSPPSRSTFN SGGGGGRTGR VPVGIGRKRL
   121  RIVVTGGAGF VGSHLVDKLI GRGDEVIVID NFFTGRKENL VHLFSNPRFE LIRHDVVEPI
   181  LLEVDQIYHL ACPASPVHYK YNPVKTIKTN VMGTLNMLGL AKRVGARFLL TSTSEVYGDP
   241  LEHPQKETYW GNVNPIGERS CYDEGKRTAE TLAMDYHRGA GVEVRIARIF NTYGPRMCLD
   301  DGRVVSNFVA QTIRKHPMTV YGDGKQTRSF QYVSDLGLVA LMENDHVGPF NLGNPGEFTM
   361  LELAEVVKEV IDPSATIEFK PNTADDPHKR KPDISKAKEQ LNWEPKISLR EGLPRMVSDF
   421  RNRILNEDEG KGL-

Predicted gene structure (within gDNA segment 154601 to 158400):

 Exon  1 154763 155338 ( 576 n);  Protein     1   207 ( 207 aa); score: 0.398
  Intron  1 155339 156123 ( 785 n);  Pd: 0.000   Pa: 0.000
 Exon  2 156124 156271 ( 148 n);  Protein   208   256 (  49 aa); score: 0.976
  Intron  2 156272 156360 (  89 n);  Pd: 0.989   Pa: 0.987
 Exon  3 156361 156440 (  80 n);  Protein   257   283 (  27 aa); score: 0.894
  Intron  3 156441 156922 ( 482 n);  Pd: 0.999   Pa: 0.791
 Exon  4 156923 157006 (  84 n);  Protein   284   311 (  28 aa); score: 1.000
  Intron  4 157007 157101 (  95 n);  Pd: 0.984   Pa: 0.469
 Exon  5 157102 157176 (  75 n);  Protein   312   335 (  24 aa); score: 0.800
  Intron  5 157177 157607 ( 431 n);  Pd: 0.268   Pa: 0.990
 Exon  6 157608 157700 (  93 n);  Protein   336   365 (  30 aa); score: 0.841
  Intron  6 157701 157772 (  72 n);  Pd: 0.977   Pa: 0.999
 Exon  7 157773 157957 ( 185 n);  Protein   366   427 (  62 aa); score: 0.803
  Intron  7 157958 158294 ( 337 n);  Pd: 0.653   Pa: 0.618
 Exon  8 158295 158312 (  18 n);  Protein   428   433 (   6 aa); score: -0.656

MATCH   AP003271+       At3g53520.1    0.652   1259    0.967   P
PGS_AP003271+_At3g53520.1      (154763  155338,156124  156271,156361  156440,
   156923  157006,157102  157176,157608  157700,157773  157957,158295  158312)

Alignment:

//

ATCCTGCTCG AGGTGGACCG GATCTATCAC CTCGCGTGCC CCGCGTCCCC TGTGCACTAC   155314
 I  L  L   E  V  D  R   I  Y  H   L  A  C   P  A  S  P   V  H  Y
 |  |  |   |  |  |  +   |  |  |   |  |  |   |  |  |  |   |  |  |
 I  L  L   E  V  D  Q   I  Y  H   L  A  C   P  A  S  P   V  H  Y       199


AAGTACAACC CCATCAAGAC GATCATATCC TTCTCGTCCC GGATCTGCAC ATACCTTTGA   155374
 K  Y  N   P  I  K  T   I
 |  |  |   |  +  |  |   |
 K  Y  N   P  V  K  T   I ...... .......... .......... ..........      207

//

TCAAATCTGG GTTTCTTAAC AATTATTACA AGACCAATGT CATGGGAACC TTGAATATGT   156154
                               K  T  N  V   M  G  T   L  N  M
                               |  |  |  |   |  |  |   |  |  |
.......... .......... ......  K  T  N  V   M  G  T   L  N  M       217


TGGGTCTGGC AAAGCGAATT GGTGCAAGGT TCTTGCTAAC TAGCACAAGT GAAGTTTATG   156214
L  G  L  A   K  R  I   G  A  R   F  L  L  T   S  T  S   E  V  Y
|  |  |  |   |  |  +   |  |  |   |  |  |  |   |  |  |   |  |  |
L  G  L  A   K  R  V   G  A  R   F  L  L  T   S  T  S   E  V  Y       237

//
```

Figure 2B

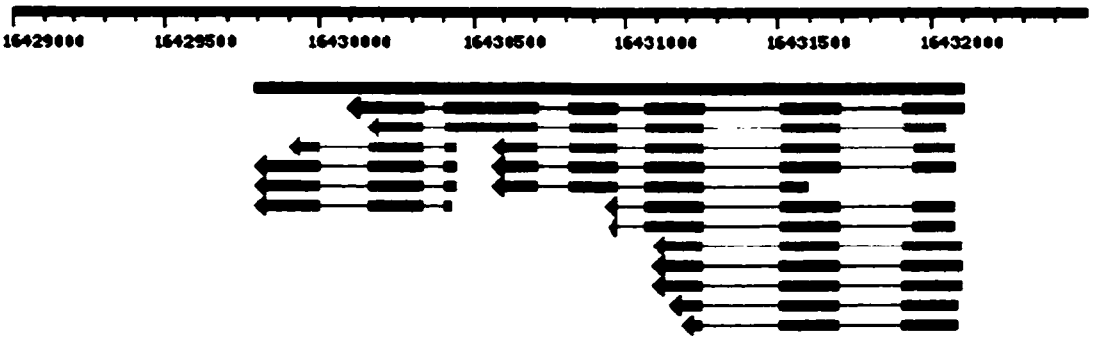

16429000
16429500
16430000
16430500
16431000
16431500
16432000

Figure 3

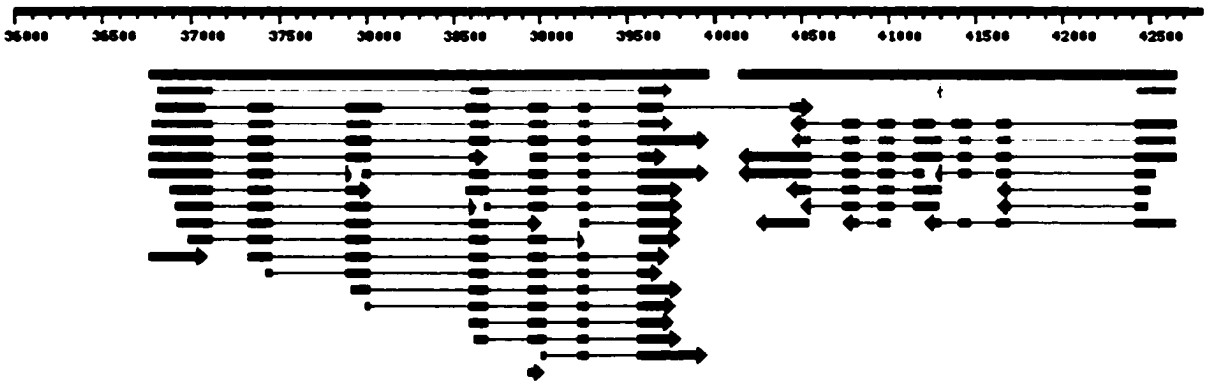

36000
36500
37000
37500
38000
38500
39000
39500
40000
40500
41000
41500
42000
42500

Figure 4

Table 1. GeneSeqer intron level performance evaluation relative to AraSet (859 annotated introns)

	Default*	Canonical Sites*	High Quality*
Predicted introns	782	684	499
Predicted introns in UTR†	76	42	28
Predicted introns in CDS	706	642	471
Correct introns	625	609	463
Overlapping introns	64	32	8‡
Wrong introns	17	1	0
Missed introns	188	235	391
Specificity	0.885	0.949	0.983
Corrected Specificity	≥0.895	≥0.961	1.000
Sensitivity	0.728	0.709	0.539

*Default, GeneSeqer default parameters; Canonical Sites, predicted canonical introns only; High Quality, canonical introns with high sequence similarity of EST to flanking exons; see Text for details

†Some of these introns are actually from unannotated genes; see Text for details.

‡Listed in Table 2.

Table 2. Annotated introns in AraSet contradicted with high quality intron predictions derived from EST spliced alignments

SeqID	Annotated Intron		Predicted Intron		EST Evidence*	Alternative Splicing
	5'ss	3'ss	5'ss	3'ss		
seq06	5753	5885	5764	5885	gi:8695314	N
seq53	3795	3708	3795	3735	gi:8715801	Y
seq62	2139	2351	2106	2351	gi:1054038	Y†
seq72	6486	6656	6481	6656	gi:4714042	Y
seq73	2232	2078	2258	2091	gi:19828992	N
seq73	3515	3398	3515	3407	gi:19868516	N
seq81	4016	3985	4088	3985	gi:14580187	N
seq84	4759	5173	4759	5170	gi:19865385	N

*Only one EST is listed for each predicted intron; for details, see Supporting Information at http://www.plantgdb.org/AtGDB/prj/ZB03PNAS/AraSet/AraSet-AtGDB.html

†See Fig. 3.

CHAPTER 3. GENE STRUCTURE IDENTIFICATION WITH MyGV USING cDNA EVIDENCE AND PROTEIN HOMOLOGS TO IMPROVE *ab initio* PREDICTIONS

A paper published in *Bioninformatics*[1]

Wei Zhu [2]and Volker Brendel[3]

ABSTRACT

Summary: MyGV is an application to visualize (potentially genome-scale) gene structure annotation and prediction. The output of any external gene prediction program can be easily converted to a generalized format for input into MyGV. The application displays all input simultaneously in graphical representation, with a toggle option for a text-based view. Zooming capabilities allow detailed comparisons for specific genome locations. The tool is particularly helpful for refinement of *ab initio* predicted gene structures by spliced alignment with cDNA or protein homologs.
Availability: The program was written in JAVA and is freely available to non-commercial users by electronic download from http://bioinformatics.iastate.edu/bioinformatics2go/MyGV.

Accurate and comprehensive genome annotation remains the foremost challenge in the post-sequencing genome era. The recent recognition of extensive alternative splicing of mammalian genes underscores the importance of the annotation task, because in most cases transcript isoforms are expected to reflect functional diversity. The theoretical foundations of gene structure are presently only partially understood. Exon prediction methods are largely based on statistical approaches. Different programs often give conflicting predictions. Large collections of Expressed Sequence Tags (ESTs) and, increasingly, full-length cDNAs can provide evidence for certain exons and thus improve *ab initio* gene prediction methods (Kan *et al.*, 2001; Gemünd *et al.*, 2001). Additionally, spliced

---

[1] Reprinted with permission of Bioinformatics, 2002 May;18(5):761-2.

[2] Primary researcher and author, graduate student, department of Zoology and Genetics, Iowa State University.

[3] Author for correspondence, professor, department of Zoology and Genetics, department of Statistics, Iowa State University

alignment with selected protein targets can often identify the gene structure of a homologous gene locus. In practice, successful gene annotation relies on careful comparison of multiple sources of prediction. MyGV was developed in response to such needs on the premise that such comparisons are most efficiently evaluated by a combination of graphical representation and analytical detail (see also Harris, 1997; Kent and Zahler, 2000; Rutherford *et al.*, 2000).

INPUT: SEQUENCE FILES AND EXTERNAL PROGRAM RESULTS

MyGV accepts as sequence input representation of a DNA molecule in common GenBank or FASTA file format. The program provides an annotation overview panel in which CDS feature entries of GenBank files are graphically represented by solid arrows extending from first to last exon and pointing in the direction of transcription. Detailed gene structure is displayed in a second, scalable view panel. Currently, no other annotation features in the sequence input files are being used. Additional input consists of formatted output of gene prediction programs, which is similarly displayed. This input is generated by piping the output of external programs (run on the same sequence input) through format converters. The current release includes format converters for Fgenesh (Salamov and Solovyev, 2000), GeneMark.hmm (Lukashin and Borodovsky, 1998), GeneSeqer (Usuka *et al.*, 2000; Usuka and Brendel, 2000), GENSCAN (Burge & Karlin, 1997), and GlimmerM (Salzberg *et al.*, 1999), but others can easily be written by the user according to need.

MyGV DISPLAY

Figure 1 illustrates the application with analysis of a segment of the *Drosophila melanogaster* genome. The input sequence file was GenBank AE002638, representing about 4.9 Mb at the terminal tip of the left arm of chromosome 2. The detailed view covers a small region including the annotated genes CG15386, CG7074, and CG7082 and represents the output of the *ab initio* programs GENSCAN (GSN), Fgenesh (FGH), and GeneMark.hmm (GM) as well as the EST spliced alignments generated by GeneSeqer (EST and AGS). The display is divided into five regions:

1. Toolbar. This section of the display controls a number of pull-down menus that are used to open and close files, execute gene prediction programs, select the zoom level, and similar functions.
2. Annotation List Tree (ALT) panel. All displayed items are listed with checkboxes that allow

selection and de-selection of individual items.

3. Annotation Overview (AO) panel. Annotated and predicted genes are represented by arrows from 5'- to 3'-extent of the coding region. Different programs are distinguished by the color scheme, e.g., GenBank, blue, GENSCAN, cyan. The vertical green lines delineate the region of the input sequence analyzed in detail in the
4. Annotation Scalable View (ASV) panel. The color scheme in the ASV panel is the same as in the AO panel, except that exon quality scores are color-coded whenever assigned by a program. Introns are shown as horizontal lines connecting the exon boxes. Vertical lines of proportional lengths flanking the introns indicate splice site scores given by GENSCAN and GeneSeqer.
5. Text Data Overview (TDO) panel. This panel tabulates details of the (predicted) exon or intron marked by a blue cross in the ASV panel. Normalized similarity and splice site scores generated by the corresponding programs are displayed whenever applicable.

The AO panel can be toggled to "text" which will display the program output of the selected item in the ASV panel. The text information can be edited in the AO panel, with an update function redrawing the graphical display in the ASV panel accordingly. This function is useful for manual refinement of the computed gene predictions. Other features of the program are described in the software documentation.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant 5R44HG01850-03. W.Z. is grateful to the Bioinformatics and Computational Biology graduate program at Iowa State University for a J. Cornette Fellowship during the first half of 2001.

REFERENCES

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78-94.

Gemünd, C., Ramu, C., Altenberg-Greulich, B. and Gibson, T.J. (2001) Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. *Nucl. Acids Res.*, **29**, 1272-1277.

Harris, N.L. (1997) Genotator: a workbench for sequence annotation. *Genome Res.*, **7**, 754-762.

Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889-900.

Kent, W.J. and Zahler, A.M. (2000) The Intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucl. Acids Res.*, **28**, 91-93.

Lukashin, A.V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucl. Acids Res.*, **26**, 1107-1115.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A. and Barrell, B. (2000) Artemis: sequence visualization and annotation.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516-522.

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24-31.

Usuka, J., Zhu, W. and Brendel, V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203-211.

Usuka, J. and Brendel, V. (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *J. Mol. Biol.*, **297**, 1075-1085.

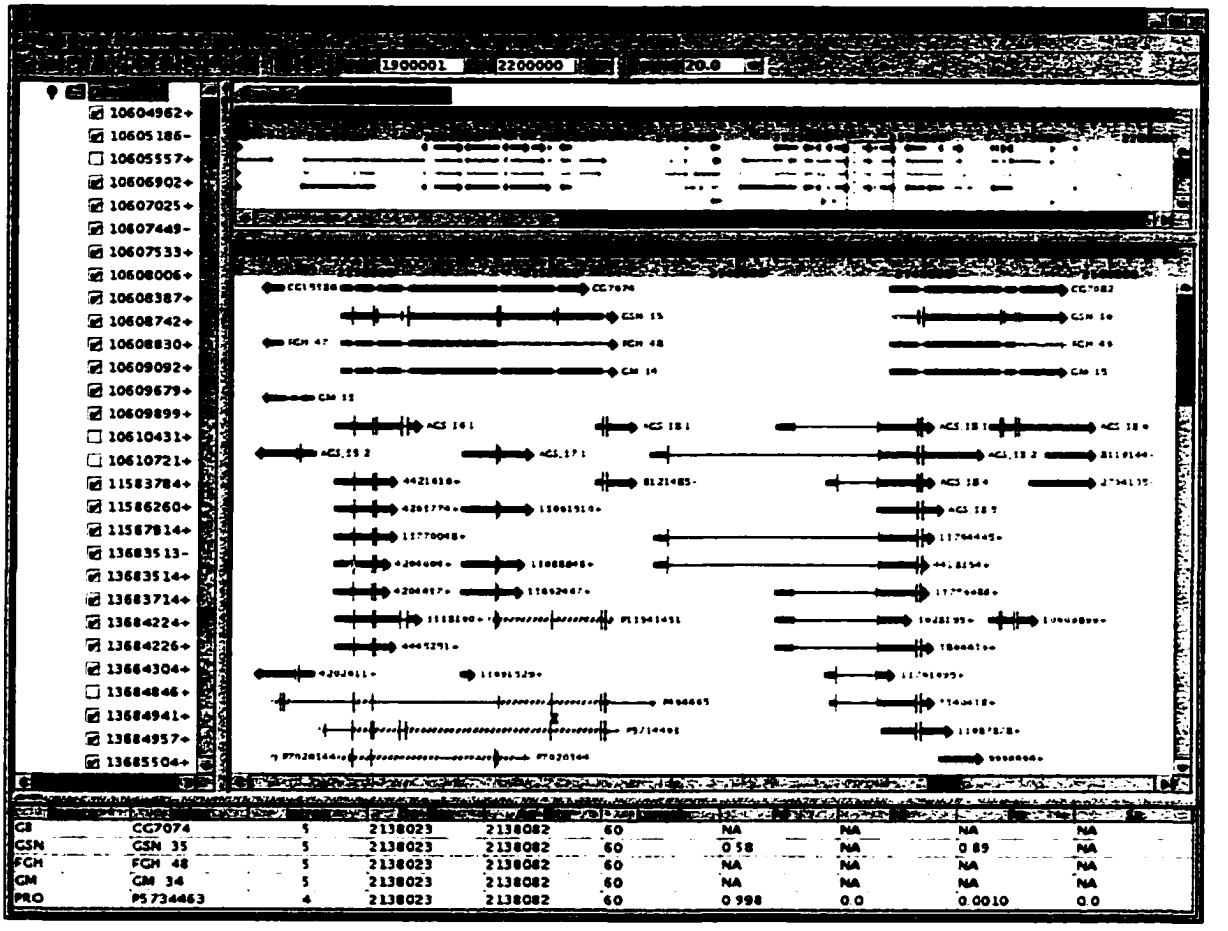


Fig. 1. Genome annotation for a segment of *D. melanogaster* chromosome 2 (GenBank AE002638) based on *ab initio* gene structure prediction programs and spliced alignment of ESTs and proteins. The five numbered MyGV display regions are described in the text. Gene prediction results are shown in panel 4. A single EST (GenBank GI:4202611) confirms intron 1 but not intron 2 of the GeneMark.hmm prediction (GM 33) upstream of the GenBank annotated CG15386 gene. GENSCAN, Fgenesh, and GeneMark.hmm give a consistent gene prediction that agrees with and extends CG7074. Introns 1-4 and 6 are confirmed by multiple EST evidence, summarized by GeneSeqer as AGS_36.1, AGS_37.1, and AGS_38.1. The blue cross in panel 4 selects intron 4 of the GeneSeqer spliced alignment with a putative *S. pombe* protein (GI:5734463) for display in panel 5. EST evidence confirms the coding exon assignments for CG7082 and supports four isoforms of the 5'-terminal untranslated region (AGS_38.2, AGS_38.3, AGS_38.4, and AGS_38.5), differing in the location of an upstream exon.

CHAPTER 4. COMPUTATIONAL MODELING OF GENE STRUCTURE IN *Arabidopsis thaliana*

A paper published in Plant Molecular Biology[1]

Volker Brendel[2] and Wei Zhu[3]

Abstract

Computational gene identification by sequence inspection remains a challenging problem. For a typical *Arabidopsis thaliana* gene with five exons, at least one of the exons is expected to have at least one of its borders predicted incorrectly by ab initio gene finding programs. More detailed analysis for individual genomic loci can often resolve the uncertainty on the basis of EST evidence or similarity to potential protein homologs. Such methods are part of the routine annotation process. However, because the EST and protein databases are constantly growing, in many cases original annotation must be re-evaluated, extended, and corrected on the basis of the latest evidence. The *Arabidopsis* Genome Initiative is undertaking this task on the whole genome scale via its participating genome centers. The current *Arabidopsis* genome annotation provides an excellent starting point for assessing the protein repertoire of a flowering plant. More accurate whole genome annotation will require the combination of high-throughput and individual gene experimental approaches and computational methods. The purpose of this article is to discuss tools available to an individual researcher to evaluate gene structure prediction for a particular locus.

---

[1] Reprinted with permission of *Plant Molecular Biology*. 2002. 48, 49-58.

[2] Primary author, and author for correspondence, professor, department of Zoology and Genetics, department of Statistics, Iowa State University.

[3] Author involved in data collection, data analysis, data interpretation, graduate student, department of Zoology and Genetics, Iowa State University.

Introduction

Modern DNA sequencing technology has revolutionized genetic research. Not long ago, the classical approach of isolating and characterizing a particular mutant would have reached a climax in the cloning and sequencing of the affected gene. Individual groups of researchers would contribute to our overall understanding of an organism or more general molecular mechanisms through their detailed studies of a particular gene or set of genes. This "one gene at a time" science has now been complemented by "high-throughput" approaches that quickly generate vast amounts of data on a large number of genes or a whole genome. Sequencing of entire genomes is the primary example of this new science, typically conducted by large research centers coordinated by national and international consortia. The sequencing of the *Arabidopsis thaliana* genome was the result of one such effort, culminating with the announcement of the complete genome in December 2000 [19]. The scope of such projects necessitates industrial approaches to data accumulation and processing, relying to a large extent on robotics and computational methods. Furthermore, this industrial approach has consequences similar to the industrialization of manufacturing: the goods delivered are produced for the entire community, and the former close connection between the craftsperson and his or her products may be lost. For genome projects, those producing the sequence can, at least initially, present only a rough overview of the features of the genome because of the scale and speed of data accumulation. The detailed understanding of particular aspects of the genome will likely have to continue to rely on the "one gene at a time" studies.

The primary task of genome annotation involves identification of gene locations and precise gene structure in terms of promoter elements, transcription signals, exon/intron boundaries, and the translation product (or possibly multiple products in case of alternative transcription start or pre-mRNA processing sites). In the context of the discussion above, the annotation task can be seen as involving two stages. The first stage is large-scale annotation, produced as the sequencing progresses and submitted to the community along with the publication of the genome sequence. For *Arabidopsis*, a total of about 25,500 protein-coding genes have been annotated in the five chromosomes [19]. Necessarily, a large number of these annotations are tentative and refer to hypothetical proteins or putative homologs. Thus, the second stage of annotation involves successive re-evaluation, extension, and correction of the annotation, removing many tentative assignments on the basis of novel experimental evidence.

The purpose of this article is to review options for the "one gene at a time" biologist who wants to use the genome information for his or her detailed studies of particular genes. In this case, he or she

cannot rely solely on the supplied genome annotation, which may well be incomplete or outdated. Instead, one must evaluate the sequences from scratch, using all particular information currently on hand, as, for example, EST evidence or potential protein homologs. We first review the principles of three prominent *ab initio* gene prediction programs for *Arabidopsis*, then discuss similarity-based prediction methods ("spliced alignment"), and lastly elaborate specific examples of evaluation of particular loci. The computational resources discussed in this article are summarized in Table 1.

Ab initio algorithms for gene finding

A large number of gene finding algorithms have been developed that produce species-specific gene structure predictions on genomic DNA without explicit comparisons to cDNAs or protein sequences. The success of these methods depends on the applicability of extrapolation of sequence features gleaned from prior training on known gene structures. The principles of many such programs are eloquently reviewed in [7]. Recently, Pavy et al. [15] evaluated programs in common use for *Arabidopsis* genome annotation and found GeneMark.hmm [13] to be the most accurate program. Also in wide use are GENSCAN [5] and GlimmerM [17]. All three programs are based on hidden Markov models. GENSCAN is built as an explicit state duration hidden Markov model. The algorithm explicitly scores for transcriptional and translational signals. Sequence composition is modeled by fifth-order Markov models, fitted according to exon phase and average C+G composition. GeneMark.hmm implements a similar model, although the details have not been described. GlimmerM uses dynamic programming to determine high-scoring combinations of coding exons. Exon/intron boundaries are determined from species-specific second-order Markov chain models, and exons are scored by fitting 3-periodic interpolated Markov models. On a large test set of validated multi-gene contigs, Pavy et al. [15] reported exon level sensitivity and specificity of about 0.8 with the best ab initio programs. A common approach for whole genome annotation is to increase the reliability of prediction by using the consensus prediction of a number of gene prediction algorithms. The combination of GeneMark.hmm, GENSCAN, and MZEF [23] led to 97% exon level specificity on the Pavy et al. set, however, with sensitivity down to 33% [15]. At the whole gene level, predicted models were found more often wrong than correct [15]. The main problem occurred with correct prediction of the proper gene boundaries. On balance, the ab intio programs are highly successful with respect to an initial annotation that can serve as a starting point for refined analysis using methods discussed in the next section, but such additional analysis remains necessary if whole gene level annotation accuracy is required.

Spliced alignment

Currently the most successful and direct method for gene identification in genomic DNA relies on cDNA sequencing with subsequent sequence alignment to the corresponding genomic DNA region. Because complete cDNA sequencing can be time-consuming and costly, high-throughput EST (Expressed Sequence Tag) sequencing has become the practical alternative to whole genome sequencing efforts. The publicly available EST collections (GenBank dbEST, http://www.ncbi.nlm.nih.gov/dbEST/) range in size from over 3.5 million entries for human to several thousands for more than 40 other species. Efficient data mining of this resource requires fast and accurate algorithms to screen an appropriate EST collection for matches against a query genomic DNA input.

The alignment of ESTs (or complete cDNAs) to eukaryotic genomic DNA typically involves long gaps corresponding to the intervening sequences that are spliced from the pre-mRNA transcript. In the absence of sequencing errors, alignment of a cognate EST to its genomic DNA source is straightforward, and a general alignment tool such as BLASTN [1] would suffice in principle. Because EST sequences are generally less reliable, specialized algorithms also take into account consensus splice site sequences to identify introns correctly even in the presence of mismatches and insertions/deletions in the alignment. The sim4 program [8] implements an efficient algorithm for such alignments under the restriction of gap-free matching in presumed exons. Introns are identified by adjusting the ends of consecutive "exon cores" (consistently ordered, close, high-scoring gap-free alignment blocks) to match the consensus 5'- and 3'-splice site signals GT and AG, respectively (or the complementary dinucleotides CT and AC).

The recent GeneSeqer algorithm [20] implements a full dynamic programming approach to derive the optimal score and spliced alignment. The within-exon alignment may contain insertions and deletions, and potential splice sites are differentially scored according to independent splice site prediction methods. Consideration of predicted splice site strength was shown to improve the performance of the algorithm in the case of imperfect sequence matching (as a result of sequencing errors or alignment of non-cognate, but homologous ESTs). The power of such "spliced alignment" with protein (rather than cDNA) targets was first demonstrated by Gelfand *et al.* with their PROCRUSTES program [9] and by Huang *et al.* with their AAT software [10, 11]. The GeneSeqer algorithm was also extended to alignment of protein sequences with genomic DNA by maximizing similarity of the inferred translation product with the target protein [21].

Case studies

The individual *Arabidopsis* researcher interested in a particular gene or gene family has unprecedented resources because of the completed sequencing of the *Arabidopsis* genome. In principle, each gene can now be uniquely identified on the chromosomes and studied in its genomic context. Because the genome annotation is as yet incomplete, the initial part of such individual research essentially involves re-annotation of the particular loci of interest. The published database annotation will provide a good starting point, but it may not have been updated since the database entry was originally submitted and thus it may be outdated or incomplete. The current *ab initio* gene prediction programs provide a second resource for such re-annotation. But if one is interested in particular loci, knowing that the average exon prediction accuracy of these programs is approximately 80% is of little comfort. For a five-exon predicted gene structure, one may suspect that one of the exons is incorrectly predicted – but which one? Or maybe this particular prediction is above or below average accuracy. Thus, as a third resource, one must look at the latest evidence provided by more recently submitted matching ESTs or potential protein homologs that may not have been available at the time that the original annotation was performed. This additional evidence may not always solve the entire annotation problem, but may at least substantiate or refute some of the predicted exons.

We discuss three typical examples drawn from the very well annotated 1.9Mb A. thaliana chromosome 4 region originally described in [3] (coordinates 7.0 to 8.9 Mb on the chromosome). The examples illustrate several possibilities that arise when comparing the given annotation (in this case, the existing but out-dated GenBank annotation) or the *ab initio* predicted gene annotation with evidence from spliced threading. The alignment of one or several more recent ESTs may provide evidence for the correctness of the given gene annotation, it may suggest re-assignment of exon and intron boundaries, or it may indicate a novel gene annotation in a previously not annotated region. The examples argue for ongoing annotation efforts that reflect current resources, including better annotation tools, vastly increased EST collections, and larger protein repositories.

New EST evidence confirms the original gene annotation

Figure 1 gives an example of supporting EST evidence displayed by the **ISUgv** genome annotation viewer (Zhu and Brendel, unpublished). The example derives from the 130-137 kb region of GenBank locus ATFCA5 (accession Z97340). The GenBank annotation according to Bevan *et al.* [3] indicates two genes in this region, dl4125c and dl4130c. The aggregate of seven overlapping

ESTs confirms the dl4125c exon/intron assignments. Interestingly, the GeneSeqer alignment for EST GenBank index (GI) 2597507 predicts the third intron (133,463 to 133,335) on the basis of a short, weakly matching 3'-most exon segment (133,334 to 133,318). In this case, the strong acceptor site score at 133,335 (score 0.94 on a scale of 0 to 1) drives the optimal alignment to this solution, and the 10-nucleotide overlap with the central ESTs GI:5841742 and GI:1216928 results in the consensus gene prediction consistent with the dl4125c annotation. In contrast, the predictions from both GENSCAN and GeneMark.hmm additionally combine several exons of the upstream dl4130c annotated gene with dl4125c into a single gene prediction (the GENSCAN gene model also extends considerably in the 3'-direction with five additional exons up to position 126,113). No ESTs match dl4130c, and no protein homologs map to this region. It is possible that all matching ESTs derive from the 3'-end of a long transcript originating in the dl4130c region. Alternatively, the lack of ESTs for dl4130c may reflect the low abundance of distinct transcripts from a second gene. Without such extra evidence, one cannot distinguish the possibilities for the N-terminal exon assignments. Compared to GeneMark.hmm and GENSCAN, GlimmerM appears to optimize for smaller gene models. Here, the GlimmerM model conformed to the downstream six exons of dl4125c, but failed to identify the upstream exons revealed by EST GI:2597507.

New EST evidence is in conflict with earlier gene annotation

A second case is displayed in Figure 2. EST evidence in the 190-200 kb region of GenBank locus ATFCA0 (accession Z97335) suggests a gene structure quite different from the original GenBank annotation, but confirms introns 1 and 6-9 of the GeneMark.hmm prediction. There are three ESTs (GIs 8698471, 8682984, 8695751) that contradict the prediction of the third intron of the GeneMark.hmm gene structure. All of these ESTs give perfect alignment over their entire length (intron-flanking alignment displayed in the upper panel in Figure 2) and match uniquely to this location in the genome. Open reading frames are stopped in all three frames in the upstream exon for the predicted direction of transcription. Thus, a likely interpretation is that these ESTs correspond to the 3'-end of a transcript and that the predicted intron is in the 3' untranslated region of such transcript. Because the *ab initio* gene prediction programs predict coding exons only, this intron could not have been predicted by any of these programs. On the basis of the EST evidence, we consider the GeneMark.hmm prediction of exons 1-3 most likely correct, with the exception of the GeneMark.hmm predicted 3'-end of the third exon, which should be replaced by the assignment given by the EST alignment. Note that EST GI:8689419 supports the GeneMark.hmm and GlimmerM

annotated start codon (perfect matching extending 17 bases upstream of the ATG) and contradicts the GenBank annotation and GENSCAN prediction. Interestingly, ESTs GI:8721769 (sampled from root tissue) and GI:9786549 (sampled from developing seed) are in conflict with respect to the first intron assignment. It is possible that the seed EST reflects inefficient or alternative splicing of the transcript.

The second gene in this region is supported by a single EST (GI:935155). A BLASTX database search revealed similarity of the EST-derived translation product to the *Arabidopsis* 22-kilodalton peroxisomal membrane protein GI:11282649, encoded at about 2.2 Mb on chromosome 4. Spliced alignment of this protein sequence with the genomic DNA identifies this locus as a homolog. The protein sequence alignment is shown in Figure 3. Both proteins have seven exons, intron positions are conserved, and strong similarity extends over all exons. Compared to this standard, the GlimmerM model correctly predicts exons 1-5 and 7, misses exon 6, and predicts an extra exon in intron 3.

This example demonstrates how the latest available evidence must be considered to give a reliable annotation. The derived annotation of two genes, one encoding a peroxisomal protein and the other a protein of unknown function, is much different from the GenBank annotation, citing a hypothetical protein of 12 exons with weak similarity to mouse laminin chain B1 precursor extending from coordinates 199,892 to 191,737. Correct and wrong annotations both lead to entries in the public protein databases. Because the protein databases are in turn used for gene prediction, the urgent need for more accurate database annotation is clear. A conservative approach, adopted by many genome centers, is to use only experimentally proven gene products for genome annotation based on similarity. However, this approach may be too conservative because similarity on the peptide level between two inferred translation products predicted from different loci is most parsimoniously explained as resulting from correct prediction of two members of a gene family (see [4] and [20] for examples and further discussion). In fact, gene structure prediction based on assignment of conserved regions as exons and variable regions as introns in comparisons of genomic DNA from distantly related but syntenic plant species may be the most powerful method for identifying unknown genes [2].

New EST evidence leads to novel gene annotation

Figure 4 gives an example of gene discovery by ESTs. Four clusters of ESTs match significantly in an unannotated region of GenBank locus ATFCA5. GENSCAN and GeneMark.hmm both predict

one gene in this region, GlimmerM predicts five. Figure 5 shows the EST alignments in the 99-104 kb region displayed by the GeneSeqer web server. A convenient feature of this interface is that the EST-predicted consensus gene structures are scanned for long open reading frames and the corresponding peptide sequences are linked as queries to NCBI BLASTP. In this example, a 180 amino acid predicted protein fragment showed strong similarity to importin alpha proteins from a number of different animal and plant species. The spliced alignment of the *Arabidopsis* chromosome 3 encoded importin alpha (GI:3122288; chromosomal coordinates 2,120,569 to 2,123,844) is shown in Figure 4 (To complicate matters further, GI:3122288 was derived from a cDNA with several differences to the chromosomal sequence. Translation of the genomic DNA results in a translation stop at the end of the penultimate exon, consistent with sequences of importin alpha proteins from tomato, *Drosophila*, and mouse). This alignment was initially puzzling because it suggests extension of the open reading frame beyond the N-terminal stop indicated in Figure 5. Closer sequence inspection resolved this puzzle as resulting from a likely error in the genomic sequence: all four ESTs GI:2733839, GI:9788101, GI:8721283, and GI:7613097 match perfectly to the genomic DNA except for a single nucleotide insertion of a G at position 102,360 in the ATFCA5 sequence. The insertion leads to the frameshift that shortens the open reading frame. This example illustrates the additional power of spliced alignment algorithms that do not require continuous open reading frames and thus can detect frameshift errors or polymorphisms. At the predicted 3'-end of the gene, the five strongly matching ESTs split into two groups of two and three ESTs. The second group appears to define an additional intron in the 3'-untranslated region for some of the transcripts of this gene.

A powerful feature of the GeneSeqer spliced alignment method is that the concurrent optimization for sequence similarity and splice site scores allows effective use of heterologous ESTs in gene structure prediction. Here, ESTs GI:935669, GI:906859, and GI:8725149 derive from the paralogous importin alpha gene on chromosome 3, yet predict four introns consistent with the cognate ESTs.

Perspective

In their recent careful evaluation of gene prediction programs for *Arabidopsis*, Pavy *et al.* [15] showed that even the best method, GeneMark.hmm [13], found the correct gene model in only 67 of 168 known genes analyzed. Prediction of mammalian gene structure appears similarly challenging [16]. These studies strongly suggest that our theoretical understanding of both transcription and RNA-processing signals remains incomplete. Predictions based on the consensus of several different

methods increases the specificity of the predictions but at the cost of much reduced sensitivity [15]. The fact that different programs perform better or worse for particular genes indicates that the current models for gene prediction are too general and might be improved if the models were trained on specific subsets of genes. Some improvement was in fact observed for *Arabidopsis* after separating two classes of genes on the basis of codon usage [14].

Here we have demonstrated by examination of a number of typical examples that additional analysis for a particular locus may significantly increase the odds of correct gene prediction relative to the average performance of *ab initio* gene prediction methods. In particular, spliced alignment with ESTs or potential protein homologs can provide substantial evidence in favor of one or another exon/intron assignment. Current methods for mammalian genome annotation seek to automate some of these additional analyses [12, 22]. Driven by these needs, genome annotation facilitates a transition of modern molecular biology. Increasingly, high-throughput and individual gene experimental approaches as well as computational methods converge to increase our detailed understanding of complex biological processes. Within the next quarter century, we anticipate an interplay of theoretical and experimental research in biology similar to the synergistic pursuit of theoretical and experimental physics in the 20th century. For a recent example, Shoemaker *et al.* [18] used microarray technology to experimentally validate and refine computational gene predictions for human chromosome 22. Similar steps for better gene prediction in *Arabidopsis* are reviewed elsewhere [6].

With continuing increases in DNA sequencing capacities, much insight may be expected from comparative sequence analysis. Studies of genomic microcolinearity in plants that have diverged over five million years or more suggests that only genic regions are highly conserved, thus providing another means of identifying genes [2]. The next generation of biologists will be well trained in bioinformatics as well as genomics approaches and be able to view biological problems from a much wider, multifaceted perspective. Such expanded view will constitute a much better approximation to biological reality than afforded within current paradigms.

Acknowledgements

V.B. was supported in part by NSF grant DBI-9872657. W.Z. was supported by a J. Cornette Fellowship from the Bioinformatics and Computational Biology graduate program at Iowa State University. The authors wish to thank Virginia Walbot for critical comments on the manuscript.

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res., 25: 3389-3402.
2. Bennetzen, J.L. 2000. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. The Plant Cell, 12: 1021-1029.
3. Bevan, M. *et al.* 1998. Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature 391: 485-488.
4. Brendel, V. and Kleffe, J. 1998. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucl. Acids Res., 26: 4749-4757.
5. Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268: 78-94.
6. Cho, Y. and Walbot, V. 2001. Computational methods for gene annotation: the *Arabidopsis* genome. Curr. Op. Biotechn. 12: 126-130.
7. Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. Hum. Mol. Genet., 6: 1735-1744.
8. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res., 8: 967-974.
9. Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. 1996. Gene recognition via spliced sequence alignment. Proc. Natl. Acad. Sci. U.S.A., 93: 9061-9066.
10. Huang, X., Adams, M.D., Zhou, H. and Kerlavage, A.R. 1997. A tool for analyzing and annotating genomic sequences. Genomics, 46: 37-45.
11. Huang, X. and Zhang, J. 1996. Methods for comparing a DNA sequence with a protein sequence. Comput. Appl. Biosci., 12: 497-506.
12. Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res., 11: 889-900.
13. Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. Nucl. Acids Res., 26: 1107-1115.
14. Mathé, C., Déhais, P., Pavy, N., Rombauts, S., Van Montagu, M. and Rouzé, P. 2000. Gene prediction and gene classes in *Arabidopsis thaliana*. J Biotechn., 78: 293-299.

15. Pavy, N., Rombauts, S., Déhais, P., Mathé, C., Ramana, D.V.V., Leroy, P. and Rouzé, P. 1999. Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. Bioinformatics 15: 887-899.
16. Rogic, S., Mackworth, A.K. and Ouellette, F.B.F. 2001. Evaluation of gene-finding programs on mammalian sequences. Genome Res., 2001: 817-832.
17. Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J., and Tettelin, H. 1999. Interpolated Markov models for eukaryotic gene finding. Genomics, 59: 24-31.
18. Shoemaker, D.D. et al. 2001. Experimental annotation of the human genome using microarray technology. Nature, 409: 922-927.
19. The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408: 796-813.
20. Usuka, J., Zhu, W. and Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16: 203-211.
21. Usuka, J. and Brendel, V. 2000. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: Increased accuracy by differential splice site scoring. J. Mol. Biol. 297: 1075-1085.
22. Yeh, R.-F., Lim, L.P. and Burge, C.B. 2001. Computational inference of homologous gene structures in the human genome. Genome Res., 11: 803-816.
23. Zhang, M.Q. 1998. Identification of protein coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis. Plant Mol. Biol., 37: 803-806.

Figure Legends

Figure 1. Genome annotation for a segment of *A. thaliana* chromosome 4 (GenBank LOCUS ATFCA5, accession Z97340) based on *ab initio* gene structure prediction programs and spliced alignment of ESTs. Results are displayed with ISUgv, a Java tool for visualization of gene structure annotation and prediction (Zhu and Brendel, unpublished). The display is divided into five regions: 1, Toolbar. 2, Annotation List Tree (ALT) panel. The checked boxes correspond to the GenBank GI identifiers of aligned ESTs. A "+" following the GI identifier indicates alignment of the strand corresponding to the GenBank entry, whereas a "-" indicates alignment of the complementary strand. AGS, Alternative Gene Structures, represent the consensus of overlapping ESTs, after removal of more tentative exon predictions. Details will be presented elsewhere. 3, Annotation Overview (AO) panel. Annotated and predicted genes are represented by arrows from 5'- to 3'-extent of the coding region. Color scheme: GenBank (GB), blue; GENSCAN (GSN), cyan; GlimmerM (GLM), pink; GeneMark.hmm (GM), gray. The vertical green lines delineate the region of the input sequence analyzed in detail in the Annotation Scalable View (ASV) panel. 4. The color scheme in the ASV panel is the same as in the AO panel, except that exon quality scores assigned by GENSCAN and GeneSeqer are color-coded. For both programs, the quality scores are normalized to a maximal value of 1.0. Exon are represented by colored boxes as follows: red, score > 0.9; pink, score > 0.8; cyan, score > 0.7; light gray, score > 0.6; gray, otherwise. Introns are shown as horizontal lines connecting the exon boxes. Splice site scores given by GENSCAN and GeneSeqer are indicated by vertical lines of proportional lengths flanking the introns. 5, Text Data Overview (TDO) panel. This panel tabulates details of the (predicted) exon or intron marked by the blue cross in the ASV panel. Pd, donor site score. Sd, similarity score for donor site flanking 50 nucleotide exon region. Pa, acceptor site score. Sa, similarity score for acceptor site flanking 50 nucleotide exon region.

The evidence of seven overlapping EST spliced alignment supports the GenBank annotation for dl4125c. The EST-derived annotation agrees with the GeneMark.hmm exon assignments in this region, but the GeneMark.hmm prediction extends 5' into the dl14130c region.

Figure 2. EST and protein spliced alignment contradict GenBank annotation. The displayed region corresponds to 191-200 kb of GenBank LOCUS ATFCA0 (accesion Z97335). Symbols are as in Fig.1. The AOV panel is toggled to display text corresponding to the alignment in the region selected by the blue cross in the ASV panel. The alignment is supported by three different ESTs. Neither

GenBank annotation nor any of the three *ab initio* programs predict the displayed intron (GENSCAN predicts the donor site but not the acceptor site). Further analysis suggests two genes in this region, one encoding a peroxisomal protein homologous to the pmb22 peroxisomal protein (GenBank GI:11282649), and the second in the downstream region encoding a protein of unknown function; see text for discussion.

Figure 3. Protein sequence alignment of the *A. thaliana* 22-kilodalton peroxisomal membrane protein (pmb22, GI:11282649, 2.2 Mb region of chromosome 4) with the predicted protein in the 194 kb region of ATFCA0 (Figure 2). Intron positions are indicated by "=". Identical residues are on black background, and conservative substitutions are on gray background.

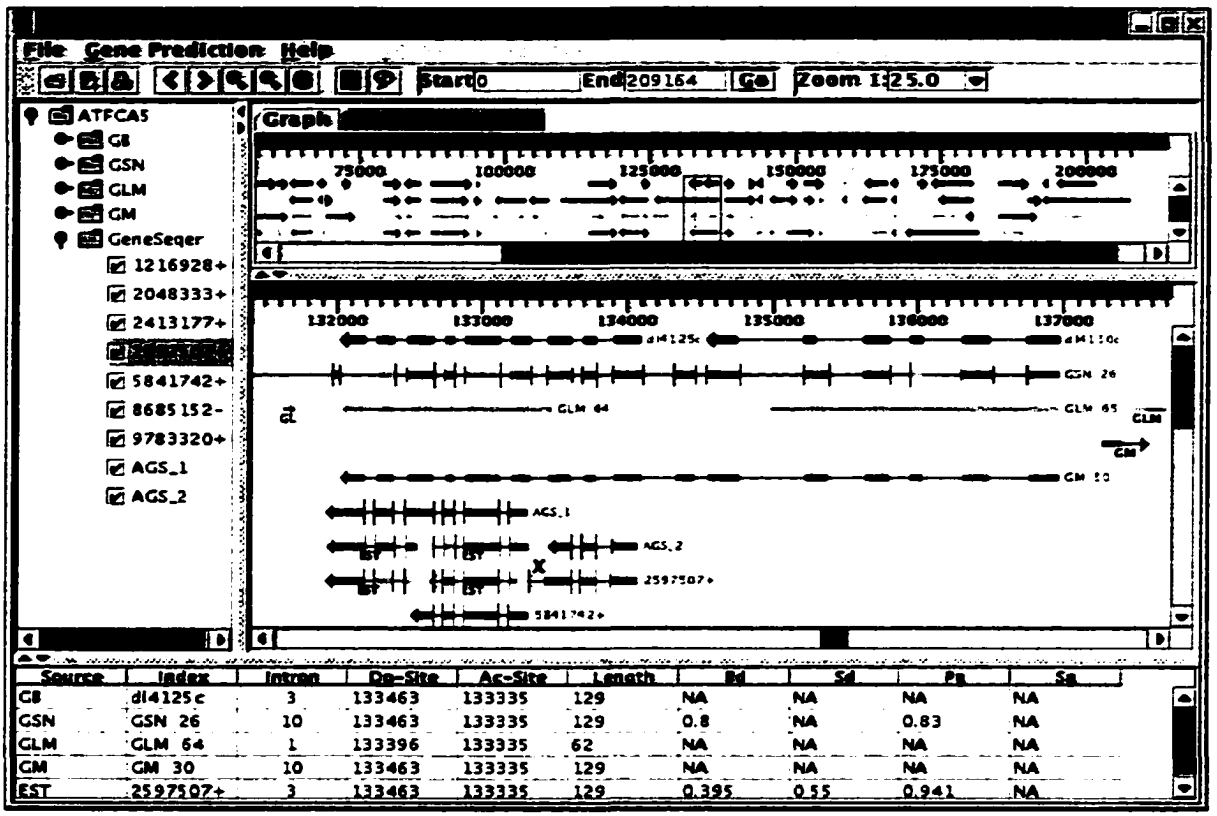
Figure 4. Gene discovery by ESTs. Two EST clusters align with the *A. thaliana* BAC GenBank ATFCA5 in the 100-104 kb region. Spliced alignment with the importin α-1 subunit (GenBank GI:3122288) suggests a 10-exon gene structure consistent with the EST evidence. Details of the alignments are discussed in the text.

Figure 5. Application of the GeneSeqer web service. The server returns the EST alignments (upper panel, blue) that are displayed in more detail in Fig. 4. The consensus gene structure prediction (green) allows two long open reading frames (red) in the 100-103 kb region. The corresponding translation product is shown in the lower panel. A BLASTP query with predicted protein fragments revealed the similarity to the importin-alpha protein that resulted in the gene prediction shown in Fig. 4.

Table 1. Some resources for computational gene structure prediction in *Arabidopsis thaliana*

Program	Web site	Reference
Ab inito *prediction*:		
GeneMark.hmm	http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi	13
GENSCAN	http:// genes.mit.edu/GENSCAN.html	5
GlimmerM	http://www.tigr.org/tdb/glimmerm/glmr_form.html	17
Spliced alignment:		
GeneSeqer	http://bioinformatics.iastate.edu/bioinformatics2go/gs.cgi	20, 21
NAP	http://bioinformatics.iastate.edu/aat/aat.html	10, 11
PROCRUSTES	http://www-hto.usc.edu/software/procrustes/qpn.html	9
Sim4	http://globin.cse.psu.edu/globin/html/docs/sim4.html	8

Figure 1.



Source	Index	Intron	Do-Site	Ac-Site	Length	Bd	Sd	Pa	Sa
GB	dl4125c	3	133463	133335	129	NA	NA	NA	NA
GSN	GSN 26	10	133463	133335	129	0.8	NA	0.83	NA
GLM	GLM 64	1	133396	133335	62	NA	NA	NA	NA
GM	GM 30	10	133463	133335	129	NA	NA	NA	NA
EST	2597507+	3	133463	133335	129	0.395	0.55	0.941	NA

Figure 2.

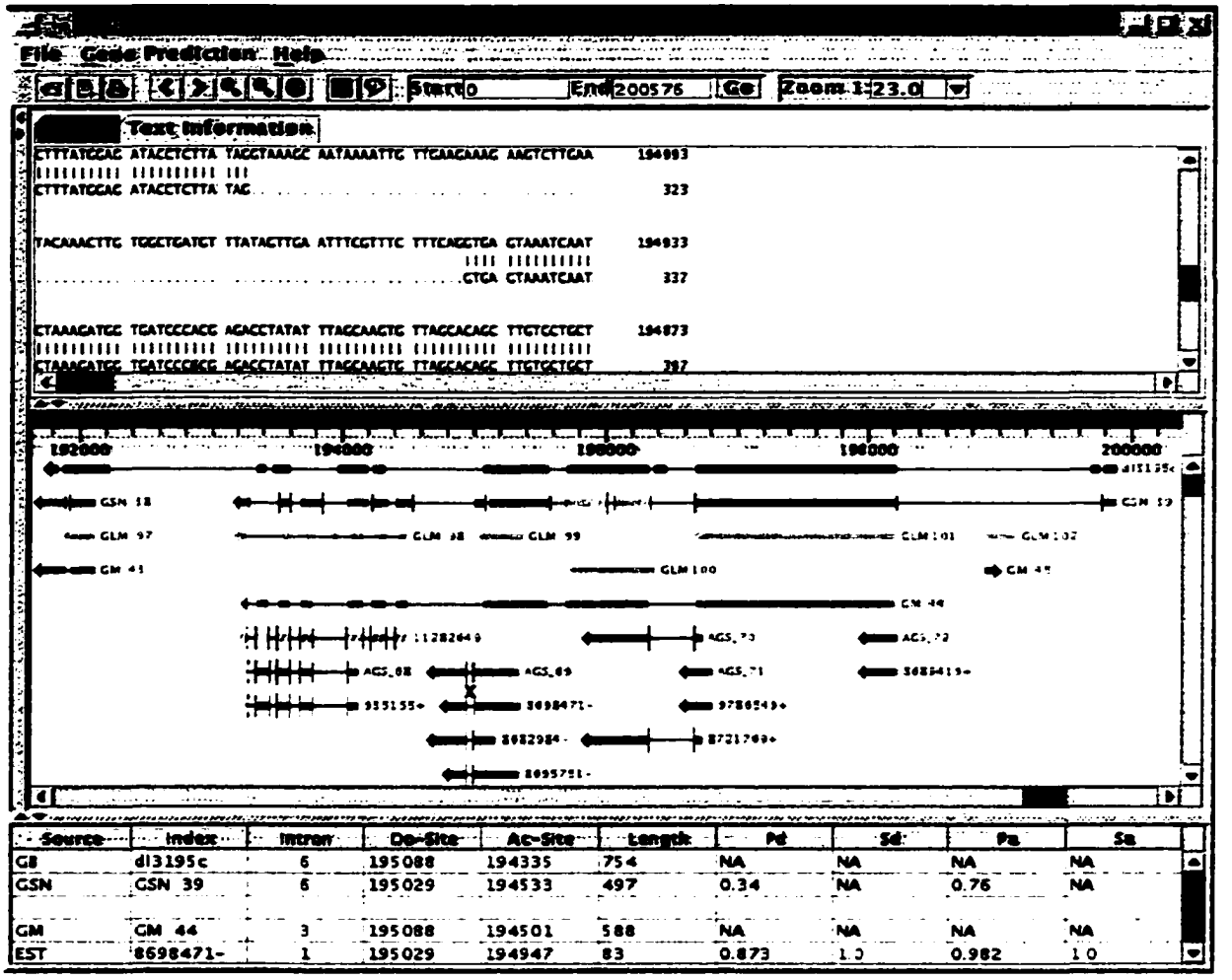

File Gene Prediction Help
Start 0
End 200576
Go
Zoom 1:23.0
Text Information
192000
194000
196000
198000
200000
GSN 18
GSN 39
GLM 97
GLM 98
GLM 99
GLM 101
GLM 102
GM 43
GLM 100
GM 44
AGS_68
AGS_69
AGS_71
Source
Index
Intron
De-Site
Ac-Site
Length
Pd
Sd
Pa
Sa
GB
dl3195c
6
195088
194335
754
NA
NA
NA
NA
GSN
GSN 39
6
195029
194533
497
0.34
NA
0.76
NA
GM
GM 44
3
195088
194501
588
NA
NA
NA
NA
EST
8698471-
1
195029
194947
83
0.873
1.0
0.982
1.0

Figure 3.

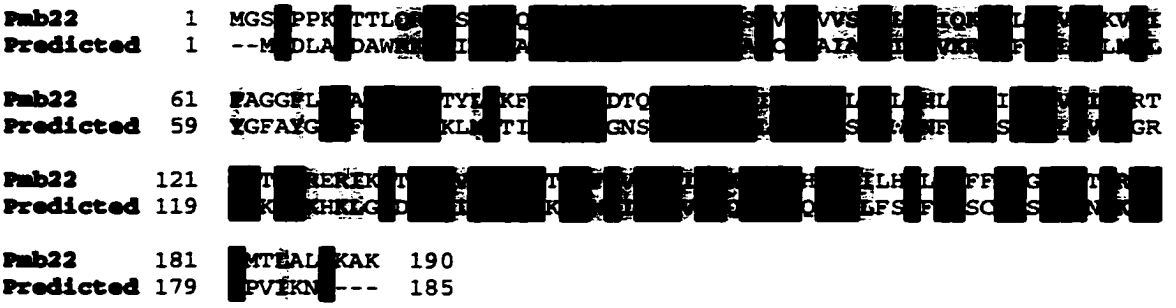

Pmb22 1
Predicted 1
Pmb22 61
Predicted 59
Pmb22 121
Predicted 119
Pmb22 181 190
Predicted 179 185

Figure 4.

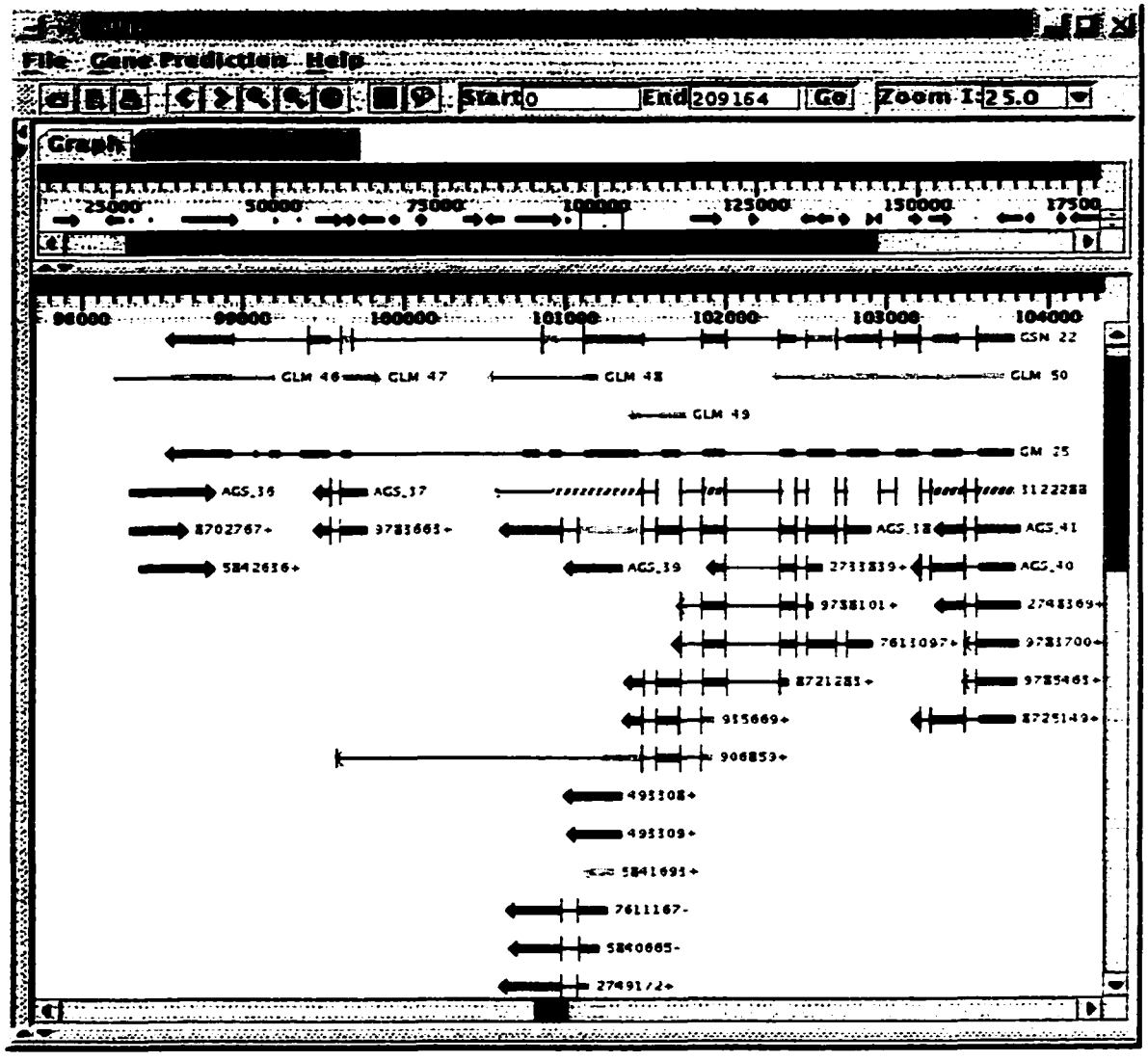
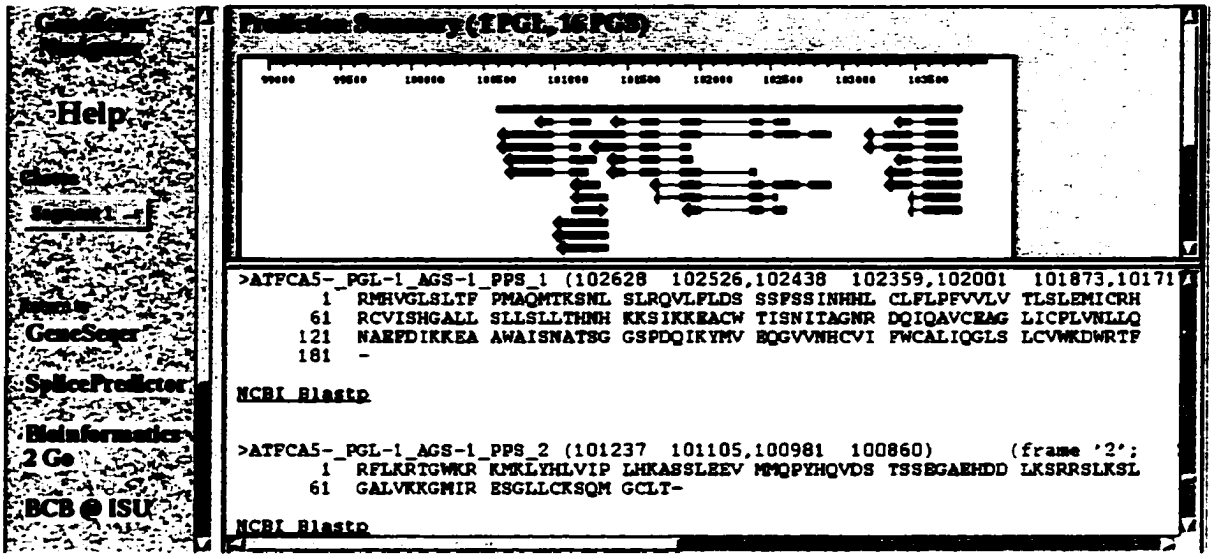

File Gene Prediction Help
Start 0
End 209164
Go
Zoom In 25.0
Graph
25000
50000
75000
100000
125000
150000
175000
98000
99000
100000
101000
102000
103000
104000
GSN 22
GLM 46
GLM 47
GLM 48
GLM 50
GLM 49
GM 25
AGS_36
AGS_37
AGS_38
AGS_41
AGS_39
AGS_40
9738101+
2748369+
7611097+
9731700+
8721281+
9735461+
915669+
8725149+
906853+
493308+
493309+
5841691+
7611167-
5840665-

Figure 5.


Help
GeneSeqer
SplicePredictor
Bioinformatics
2 Go
BCB @ ISU
>ATFCA5-_PGL-1_AGS-1_PPS_1 (102628 102526,102438 102359,102001 101873,10171
1 RMHVGLSLTF PMAQMTKSNL SLRQVLFLDS SSFSSINHHL CLFLPFVVLV TLSLEMICRH
61 RCVISHGALL SLLSLLTHNH KKSIKKEACW TISNITAGNR DQIQAVCEAG LICPLVNLLQ
121 NAEFDIKKEA AWAISNATSG GSPDQIKYMV EQGVVNHCVI FWCALIQGLS LCVWKDWRTF
181 -
NCBI Blastp
>ATFCA5-_PGL-1_AGS-1_PPS_2 (101237 101105,100981 100860) (frame '2';
1 RFLKRTGWKR KMKLYHLVIP LHKASSLEEV MMQPYHQVDS TSSEGAEHDD LKSRRSLKSL
61 GALVKKGMIR ESGLLCKSQM GCLT-
NCBI Blastp

CHAPTER 5. REFINED ANNOTATION OF THE *Arabidopsis thaliana* GENOME BY COMPLETE EST MAPPING

A paper has been accepted by *Plant Physiology*[1]

Wei Zhu[2], Shannon D. Schlueter[3], and Volker Brendel[4]

ABSTRACT

Expressed Sequence Tags (ESTs) currently encompass more entries in the public databases than any other form of sequence data. Thus, EST data sets provide a vast resource for gene identification and expression profiling. We have mapped the complete set of 176,915 publicly available *Arabidopsis* EST sequences onto the *Arabidopsis thaliana* genome using GeneSeqer, a spliced alignment program incorporating sequence similarity and splice site scoring. About 96% of the available ESTs could be properly aligned with a genomic locus, with the remaining ESTs deriving from organelle genomes and non-*Arabidopsis* sources, or displaying insufficient sequence quality for alignment. The mapping provides verified sets of EST clusters for evaluation of EST clustering programs. Analysis of the spliced alignments suggests corrections to current gene structure annotation and provides examples of alternative and non-canonical pre-mRNA splicing. All results of this study were parsed into a database and are accessible via a flexible web interface at http://www.plantgdb.org/AtGDB/.

---

[1] Accepted February 20, 2003

[2] Primary researcher and author, graduate student, Department of Zoology and Genetics, Iowa State University.

[3] Graduate student who designed, implemented the web interface of AtGDB and part of data analysis, Department of Zoology and Genetics, Iowa State University.

[4] Author for correspondence, Professor, Department of Zoology and Genetics, Department of Statistics, Iowa State University

INTRODUCTION

The efforts of an international collaboration to obtain the complete genome sequence of the flowering plant *Arabidopsis thaliana* resulted in the release and annotation of 115.4 megabases (Mb) of the genome (estimated at 125 Mb) in December of 2000 (Arabidopsis Genome Initiative, 2000). At that time, 25,498 protein-coding genes were identified in the five haploid chromosomes, but only 9% of these genes had been characterized experimentally and only 69% could be functionally classified by similarity to proteins of known functions (*ibid.*). In the interim, sequencing and annotation has progressed. The most current release of the *Arabidopsis* genome available at GenBank provides 117.3 Mb and 27,288 annotated protein-coding genes (see Data Sets in METHODS). Annotation of the *Arabidopsis* genome and functional characterization of all the genes is an ongoing effort. Initial, high-throughput computational gene structure prediction has likely been successful in identifying most gene locations; however, these methods still suffer from limitations in predicting the precise gene structure for an entire gene, detection of intergenic regions, and identification of non-coding exon sequences (Pavy et al., 1999; Brendel and Zhu, 2000). Recent studies have concentrated on sequencing of full-length cDNAs to improve genome annotation (Haas et al., 2002; Seki et al., 2002).

Expressed Sequence Tags (ESTs) are single-pass sequencing reads of cDNA clones that have become a widely employed method for gene identification, expression profiling, and polymorphism analysis. Presently, more than 13.4 million EST entries have been deposited into the NCBI dbEST public database, including *Arabidopsis* with 176,915 ESTs and 21 other species with EST sets of more than 100,000 entries (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). In the absence of a whole genome sequencing project for a particular species, clustering of ESTs into contigs that represent unique genes is one of the most promising strategies to glimpse the gene space of that organism. Challenges of EST clustering arise from poor average sequence quality, incomplete EST sampling, polymorphisms, alternative transcript isoforms, representation of highly similar transcripts from distinct members of multigene families, and cloning artifacts. Different strategies for EST clustering and the associated gene indexing databases have been reviewed by Bouck et al (1999); for a recent method for EST clustering on parallel computers see Kalyanaraman et al. (2003).

For *Arabidopsis*, up-to-date EST clusters are available in form of the UniGene clusters at NCBI (http://www.ncbi.nlm.nih.gov/UniGene/) and as a TIGR Gene Index (AtGI, http://www.tigr.org/tdb/tgi/agi/, Quackenbush et al. 2001). The current UniGene build (#28) comprises 27,248 clusters derived from 220,191 sequences (including 55,519 mRNAs). The current

AtGI (Release 9.0) comprises 38,462 clusters from 232,136 sequences. Whereas UniGene clusters are meant to represent all transcript isoforms derived from a gene locus, different transcript isoforms should split into distinct TIGR clusters. In either case, the clusters are constructed on the basis of mRNA sequence comparisons only. This is of course necessary for most species for which only limited genome sequence data are available. Here we present results of *Arabidopsis* EST clustering based on direct spliced alignment of the ESTs onto the *Arabidopsis* genome. This approach has significant advantages, when applicable (e.g., Kan et al. 2001; Yeh et al. 2001). First, accuracy should be greatly increased because cognate EST genomic locations can be easily identified for most ESTs, and clusters can be determined by proximity of EST locations on the genome scaffold. Second, the spliced alignments provide a rich data source to probe the extent and characteristics of alternative splicing, non-canonical splice sites, and other features of gene structure. We provide different sets of genome-confirmed EST clusters that can serve as standards for the comparison of programs and parameter settings for mRNA-based EST clustering. We discuss differences between current *Arabidopsis* gene structure annotation and EST-based gene annotation. All alignment results were imported into a relational database that is accessible via the web at http://www.plantgdb.org/AtGDB/ and includes extensive tools for visualization and further analysis.

RESULTS

EST Spliced Alignments

The *Arabidopsis thaliana* EST (ATest) dataset employed in this study consists of 176,915 entries. As shown in Fig. 1, only 2,059 EST sequences (1.2%) did not show any significant alignments with the genome. Further investigation based on BLASTN (Altschul et al. 1997) searches against the non-redundant nucleotide database at NCBI (E-value < 1e-10) showed that about 40% (822) of those unmatched ESTs have no hits, about one fifth (401) resulted from contamination (matching sequences from clone vectors, insects, fungi, etc.) or low complexity sequences, and another 27% (557) came from the organelle genomes (mitochondrial and chloroplast). Surprisingly, most of the remaining sequences were found to have significant hits against sequences from *Arabidopsis*. Failure of these sequences to produce a valid spliced alignment could be attributed to either of two causes. First, the matching genomic sequences have not yet been assembled into the published *Arabidopsis* genome sequence. Thus, some ESTs clearly match with *Arabidopsis* BACs (for example, EST gi:19837354

matches with BAC gi:18149207 derived from the centromere region of chromosome four) but do not match with the released *Arabidopsis* genome sequence. Second, with default parameters, GeneSeqer does not detect weak matches that may arise from poor sequence quality (for example, EST gi:9783909) or low complexity regions (for example, EST gi:9787792). Such failed alignments are expected because no repeat masking or quality clipping was performed to preprocess the EST sequences before aligning them with the genome.

96.0% of the ESTs have at least one high-quality spliced alignment (hqSPA, see METHODS) with the *Arabidopsis* genome (such ESTs denoted as hqEST), and about 13.2% have more than one hqSPA with the genome (such ESTs denoted as mhqESTs; see Fig. 1). The distribution of the number of hqSPAs per hqEST is shown in Table 1. The majority of the ESTs have only one or two hqSPAs, but there are 38 ESTs with at least 10 hqSPAs. These ESTs were found to be associated with transposon families and other highly prolific genome elements. For example, EST gi:9787698 (with 170 hqSPAs) appears to be derived from an *Arabidopsis* putative retroelement polyprotein gene, clustered around all five centromeres of the *Arabidopsis* genome as shown in Fig. 2.

Overall, about 82.8% (146,527 entries) of the ATest dataset are unique high-quality ESTs (uhqESTs, see METHODS), which align with a single locus in the genome. In order to properly position the remaining ESTs, which display multiple hqSPAs, we make the assumption that for each EST the alignment with maximal score (similarity score times coverage score) identifies the true cognate location of that EST. Such alignments are designated putative cognate spliced alignments (pcSPAs, see METHODS). In this way, 172,137 pcSPAs were generated from 169,888 hqESTs and 206,833 hqSPAs. Because of virtual equalities among the scores of some hqSPAs for certain mhqESTs (Fig. 3), there are more pcSPAs than hqESTs.

We should emphasize that our restriction on hqESTs largely eliminates typical problems of EST clustering and EST-based gene annotation, as caused for example by chimeric clones. Thus, chimeric sequences would typically lead to alignments with coverage score below 0.8, because in any given genomic location only one part of the sequence would match (or, if the foreign sequence were only very short, it would not be used in the GeneSeqer spliced alignment, which optimizes the local alignment score). According to the aforementioned assumption, the similarity and coverage scores for each pcSPA correlate with our confidence in the prediction of cognate transcript origin for the hqEST in question. Higher alignment similarity and coverage scores denote greater confidence. The vast majority of pcSPAs have similarity and coverage scores in the 0.99 to 1.0 range (Fig. 4). This implies high confidence in the classification of these alignments as cognate. The designation of

"putative" cognate is formally accurate, however, because the matched ESTs and genomic sequences were not isolated from the same plant. When considering the alignment of ESTs not derived from the Columbia ecotype on which the genomic sequences are based, cognate position implies the cognate origin of the most probable transcript ortholog to the aligned EST. According to dbEST annotation, about 98% of the *Arabidopsis* ESTs were derived from the Columbia ecotype. 300 of the 337 ESTs annotated as derived from ecotype Landsberg have pcSPAs with average similarity score 0.93 and average coverage score 0.94. Thus, the different *Arabidopsis* ecotypes appear to have such a high degree of sequence conservation that correct mapping of the ESTs onto the Columbia ecotype genome is unproblematic (see also Haas et al., 2002).

EST Clustering and Assembly

EST assembly refers to the problem of finding the correct orientation and order of EST sequences in a tiling path covering the cognate mRNA. Because EST sequences are typically generated by single-pass sequencing and thus contain a fair number of errors and ambiguous bases, this assembly can be difficult in the absence of genome sequence data. However, when the entire genome sequence is available, the spliced alignment of ESTs gives reliable assemblies and can be used for prediction of gene structure and alternative splicing (Kan et al. 2001; Yeh et al. 2001).

Because of the relative facility of EST sequencing, EST projects have outpaced genome sequencing projects for many species. EST clustering is typically the first analysis step in deriving a "unigene" set representing the transcriptome of the species. By clustering, EST sequences that share significant sequence similarity are partitioned into presumed gene-specific contigs, thus reducing the redundancy of the EST set. Such reduction is often dramatic, especially in the case of EST sets not derived from normalized libraries. Cluster-based reduction may be a practical necessity prior to EST assembly for large EST sets. Here, we are particularly interested in evaluating the utility of ESTs in gene identification. pcSPAs, representing putative cognate gene locations, were clustered based on chromosome location. Each cluster contains ESTs from a single gene provided that the intergenic regions between neighboring genes are sufficiently long compared to the maximal allowed gap (negative overlap) set by the clustering parameters (see METHODS). Because genome-based EST clustering does not depend on pairwise EST sequence overlap, which is a necessary requirement for comparison-based assembly programs, small gaps in local genome coverage can be allowed, thereby joining partial gene annotations through a genome scaffolding scheme. In addition, high coverage as

required for the pcSPAs excludes erroneous alignment of chimeric clones, which typically pose annoying problems for comparison-based assembly.

Fig. 5 provides an example of the possibilities and difficulties of gene structure annotation by EST clustering. Full-length cDNA evidence indicates four genes in alternating directions in the displayed region of chromosome four. Current GenBank annotation misses the second gene, the 5'-end of which is overlapping the 5'-end of the third gene transcribed on the opposite strand. Genome-based EST clustering without using clone pair information would give the three clusters that are bounded by ESTs gi:19864852 & gi:19802435, gi:19822861 & gi:19863255, and gi:8732113 & gi:19863376, respectively. If clone pair information is used, the clusters resolve to four clusters that correctly identify the four genes. For comparison, Fig. 5 also shows the alignment of TIGR Arabidopsis Gene Index tentative contigs (Quackenbush et al. 2001). Note the erroneous concatenation of ESTs in TC159466 and TC160975, resulting from clustering based on significant overlap only (but not coding strand identification).

Choosing various clustering parameters from 50 bp overlap to 100 bp gap was shown to alter the number of clusters by less than 12% (Table 2). The following results are based on the 27,611 clusters obtained by allowing a maximal gap of 60 bp (other criteria give similar results; data not shown). About half of the clusters contain only one or two pcSPAs (Table 3). Large clusters correspond to highly expressed genes (e.g., Fernandes et al. 2002), including ribulose-bisphosphate carboxylase, photosystem II type I chlorophyll a/b binding protein, seed storage protein, and ribosomal proteins (for descriptions of all clusters with at least 100 pcSPAs, see http://www.plantgdb.org/AtGDB/prj/ZSB03PP/virtualNorthern.html). More than 64.5% (17,609) of the annotated genes have at least one pcSPA within their annotated boundaries, with an average of about seven ESTs supporting each of these annotations (range: 1 to 1,014). 22.5% (6,141) of the 27,288 annotated gene coding regions are fully covered by an EST cluster, and for 44.8% (12,226) of the annotated genes, clone pair joined clusters confirm the annotated extent of the coding region.

Gene Identification by ESTs

As described in the next section, some of our spliced alignment results contradict particular gene models in the most recent *A. thaliana* genome annotation. To safeguard against possible errors in our employed methods, we exploited a set of 5,000 non-redundant full-length cDNAs derived in a Ceres/TIGR collaboration (Haas et al. 2002; ATcdna, see METHODS) for benchmarking. In particular, we sought to determine, first, whether the cDNA spliced alignments were consistent with

the genome annotation and, second, how the EST spliced alignments and assemblies compared with the cDNA spliced alignments. It should be noted that the Ceres/TIGR full-length cDNAs were derived from the Wassilewskija and Landsberg *erecta*, rather than Columbia, *A. thaliana* ecotypes; however, Haas et al. (2002) reported more than 99% average identity between the three ecotypes, confirmed by our spliced alignment results.

The results showed that 4,999 of the cDNAs have at least one hqSPA. The only unmatched cDNA (gi:21405014, Ceres ID: CT23693) matches mitochondrial DNA. Generally, the pcSPA of a full-length cDNA is regarded to be the most decisive experimental evidence to define gene structures. Therefore, the cDNA-derived pcSPAs provide a reliable set to assess EST-based gene prediction. Overall, the 4,999 cDNAs have 4,691 uhqSPAs, 308 mhqSPAs, and 5,013 pcSPAs. Surprisingly, 1,100 (21.9%) of the pcSPAs are embedded in longer EST clusters (see http://www.plantgdb.org/AtGDB/prj/ZSB03PP/extendedCoverage.html). This discrepancy may result from alternative transcription initiation and termination sites or systematic biases in the cDNA cloning process (Haas et al., 2000). Alternative transcription initiation and termination sites may also reflect polymorphisms among different A. thaliana ecotypes. 91.0% (4,563) of the pcSPAs are at least partially covered with ESTs, with an average of 10 EST-derived hqSPAs supporting each (partially) covered gene (range: 1 to 652). On the intron level, 81.8% (13,980) of 17,091 introns (including low-quality introns) deduced from pcSPAs of full-length cDNAs are supported by EST alignments. The majority of the annotated introns are consistent with the high-quality introns derived from cDNA spliced alignments as we expected, but there are still 28 annotated introns that are contradicted, associated with 23 distinct annotated genes (see http://www.plantgdb.org/AtGDB/prj/ZSB03PP/geneAnnotationVScdna.html).

Because the cDNA-covered gene set is not representative of the entire *Arabidopsis* gene set (highly expressed genes have a greater chance to be cloned and sequenced both as ESTs and full-length cDNAs), the 91% fraction of pcSPAs from full-length cDNAs covered also by the EST-derived pcSPAs is an upper bound of the estimated fraction of genes identified by ESTs. The comparison confirms that both ESTs and cDNAs were accurately mapped to the genome with our method, and that these approaches provide both alternative and complementary paths to gene discovery.

Arabidopsis thaliana Genome Database (AtGDB)

Spliced alignments of ATest and ATcdna as well as the recent annotation of the *Arabidopsis* genome were parsed and imported into a MySQL relational database, which was named *Arabidopsis thaliana* Genome Database (AtGDB). An elaborate web interface was designed for the database to allow users to browse the genome and query the database by sequence similarity, identifiers, or description (http://www.plantgdb.org/AtGDB/). In general, the web interface is composed of three parts: the genomic context view, the query view, and the sequence view. The genomic context view allows users to browse a specific genomic region in the context of multiple annotation resources. The region graphic displays these multiple sources of alignment information relative to one another. Each is colored with respect to its specific annotation source (see Fig. 5). The query view allows users to view and interact with the results of a user query. Stored EST/cDNA alignments and annotated transcripts each have an individual page, the sequence view, which glues together sequence data, analysis tools and related external links. This web interface efficiently presents the database entries on-the-fly, and facilitates data access and utilization as described below.

Applications

After mapping the ESTs to the genome, we not only acquired the genomic loci each EST originated from, but also confirmation of other annotation resources by comparison to the EST spliced alignments. Here, we explored several applications listed below. However, we should emphasize that we cannot describe in-depth analysis of these data within the scope of this manuscript and rather wish to point out possibilities of further studies based on the rich data source provided by the comprehensive EST mapping.

Consistency of gene structure annotation

The annotation of the *Arabidopsis* genome referred to in this study was published on Aug. 20, 2002, and represents the most current genome annotation released by the Arabidopsis Genome Initiative. Because much of the annotation is still computationally produced without human expert scrutiny, EST evidence may not always have been incorporated into the gene models. To estimate the extent of this problem, we compared annotated intron positions with predicted intron sequences based on our EST spliced alignments. As a result, 58,120 out of the 115,949 annotated introns were

confirmed. Another 1,272 annotated introns are inconsistent with high-quality predicted introns inferred from the spliced alignments. These introns occur in 977 distinct gene models or about 3.4% of the annotated genes (data available at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/geneAnnotationVSest.html). Although these discrepancies may be caused by alternative transcript isoforms, erroneous gene prediction seems a more parsimonious explanation in the absence of other evidence.

In addition to suggesting corrections to current gene annotations, the EST spliced alignments also identify novel gene locations. Thus, of the 27,611 EST contigs assembled on the basis of proximity in their genomic locations, 129 occur in regions without any annotated gene models and contain open reading frames longer than 100 residues that show no significant hits with annotated *Arabidopsis* proteins using BLASTP (threshold 1e-10). 82 of these show no hits at the same threshold when compared against the NCBI non-redundant (nr) protein database, and the remaining 47 EST contigs show at least one hit (data available at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/novelGenes.html). For example, ESTs gi:19863912, gi:9786135 and gi:8721866 form a cluster that supports an ORF of 108 residues between genes At4g02400 and At4g02410; the existence of a gene in that region is also supported by full-length cDNAs gi:14596167 and gi:20148266. In other cases, the novel ORFs may correspond to upstream or downstream exons of incompletely annotated genes. The display at AtGDB allows users to provide updated annotation upon more in-depth analysis of individual cases.

5′- and 3′- Untranslated Regions (UTRs) in mRNAs

Most annotated gene models correspond to the coding portions of exons only. Although attempts have been made recently to predict the UTR portions of mRNAs by genome sequence inspection (Davuluri et al. 2000, 2001; Tabaska et al. 2001), this has proven to be a difficult endeavor. If UTRs are annotated, the annotations are derived mostly from full-length cDNAs. ESTs provide a more accessible resource to gain UTR information, provided accurate EST assembly and mapping onto the genome is possible.

The gene density in the *Arabidopsis* genome is high, with about one gene every 5kb. Therefore, intergenic regions are typically very short, which may make accurate UTR assignments difficult. We cataloged high-quality predicted introns that mapped into annotated intergenic regions into potential 5′-UTR or 3′-UTR introns depending on whether the constituent hqSPAs extend from the flanking

coding region into the upstream or downstream region, respectively (note that in some cases the additional exons may extend an annotated open reading frame; thus, the derived set of *potential* UTR introns is a superset of EST confirmed UTR introns). In this way, 2,282 potential 5'-UTR introns in 2,023 annotated genes (including 199 genes with multiple potential 5'-UTR introns; all data displayed at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/upstreamUTRintrons.html) and 570 potential 3'-UTR introns in 487 annotated genes (including 47 genes with multiple potential 3'-UTR introns; all data displayed at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/downstreamUTRintrons.html) were identified. 72 genes have both potential 5'-UTR and potential 3'-UTR introns. Thus, at least 9% of *Arabidopsis* genes may have introns in their UTRs. Our listing of these features at AtGDB should provide a valuable resource to study possible roles for these introns in the regulation of gene expression and to develop models for UTR prediction (see also dbUTR, Pesole et al. 2002).

Non-canonical splice sites

Almost all introns contain the canonical GT-AG splice site junctions, but other varieties also exist. It was estimated that about 1% of *Arabidopsis* introns are non-canonical GC-AG introns (Brown et al. 1996), slightly higher than the proportion identified in mammals (Burset et al. 2000, 2001). In all other respects, GC-AG introns seem to be analogous to the canonical GT-AG introns, and they are processed in the same splicing pathway (U2-type spliceosome). AT-AC introns, with consistently low frequency in diverse eukaryotic taxa, are another well-studied type of non-canonical introns, which are typically spliced by a distinct U12-type spliceosome (Wu and Krainer, 1996; Wu et al., 1996; Burge et al., 1998).

In this study, 738 introns (1.7% of the 43,165 high-quality predicted introns derived from EST alignments) were found to have non-canonical splice sites (Table 4). GC-AG introns represent the large majority of non-canonical introns (453 cases, or about 1.0% of all high-quality predicted introns). AT-AC introns comprise the second largest category (25 cases). Many of the non-canonical introns have short direct repeats spanning the donor and acceptor sites. In these cases, the exact intron position cannot be unambiguously determined by spliced alignment, and thus some of the classifications in Table 4 may prove incorrect. The complete listing of apparent non-canonical introns (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/ncSpliceSites.html) should facilitate experimental investigation of splicing in the absence of the standard splice site features.

The 453 GC-AG introns (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/non_canonical/gc_ag.html) have the consensus donor sequence [nonU]AG/GCAAGU (donor site boldfaced) exactly as reported before for other data sets (Burset et al. 2000). These introns exhibit a similar distribution of predicted splice site scores as do GT-AG introns (Brendel and Kleffe 1998; data not shown). This suggests that the mechanism of splicing of GC-AG introns may be the same as that of GT-AG introns but involve more highly conserved sequence features apart from the GC dinucleotide.

Dietrich et al. (1997) reported that U12-type introns are more likely to be determined by the conserved motifs around the donor site and the branch site than by the dinucleotide-termini of the intron. Consistently, some AT-AC introns are spliced by the U2-type spliceosome, whereas some GT-AG introns are spliced by the U12-type spliceosome (*ibid.*). In this study, all but two of the 25 AT-AC introns (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/non_canonical/at_ac.html) exhibit both the ATATCCTY donor site motif and the TCCTTRAY branch site element (Wu and Krainer, 1996; Burge et al., 1998). The two exceptions (derived from the uhqSPAs of ESTs gi:931334 and gi:19874656) may not be typical U12-type AT-AC introns and could also be classified as non-canonical TT-CC and TC-CA introns, respectively. In addition, one AT-AA intron and 17 GT-AG introns were identified as likely U12-type introns based on a more detailed motif search (see METHODS). All 41 likely U12-introns and related information are listed at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/u12Introns.html.

Mutual comparison of the genes containing the putative U12-type introns shows that some of them may correspond to duplications within gene families. For example, the genes At1g56280, At5g26990, At3g06760, At3g05700, and At4g02200 all encode a drought-induced-19 like protein. A detailed study shows that all five genes have a U12-type intron between coding exons three and four (At4g02200 has a U12-dependent GT-AG intron, whereas the other four genes have a U12-dependent AT-AC intron). Similarly, the genes At3g53520 (Fig. 6A) and At3g62830, which encode a dTDP-glucose 4-6-dehydratase like protein, also both have an U12-dependent AT-AC intron in the same location. Inspection of a homologous rice gene shows that the U12-type intron location is not only conserved among the *Arabidopsis* paralogs but also across the monocot/dicot divide (Fig. 6B). This observation is consistent with the conjecture of the early origin of U12-class introns (Wu et al., 1996; Burge et al., 1998; Wu and Krainer, 1999).

The analysis of U12-type introns gives an example of how to utilize the EST data and AtGDB resource, and it also exposes several annotation problems. For instance, of the 23 AT-AC U12-type

introns, only four AT-AC introns are explicitly annotated (At3g53520, At5g22650, At5g26990, and At5g27380). One AT-AC U12-type intron in gene At3g62830 is incorrectly annotated as a CA-TA intron, even with the presence of six cognate full-length cDNAs. In addition, gene structures predicted by *ab initio* methods will typically never include non-canonical introns (for example, the gene At1g76170). Furthermore, EST data can provide a check on the accuracy of the genome sequence (Brendel & Zhu, 2002). For example, 24 ESTs supporting the AT-AC U12-type intron in the drought-induced-19 like gene At1g56280 clearly suggest that one adenosine should be inserted after the 20,673,745 bp position in chromosome one of the current genome assembly. This inference is also supported by two cognate full-length cDNAs.

Alternative Splicing

Current research suggests that approximately 40-60% of human genes are alternatively spliced (Black 2000; Brett et al. 2002; Modrek and Lee 2002). Identification of alternative splicing is generally based on cDNA or EST evidence (Coward et al. 2002; Huang et al. 2002; Kan et al. 2001; Modrek and Lee 2002). Based on strict criteria and manual inspection (see METHODS), we identified 327 cases of alternative splicing among *Arabidopsis* genes and categorized them into five groups: (1) alternative donor sites (102 cases), (2) alternative acceptor sites (190 cases), (3) alternative introns that are shifted in position at both sites (3 cases), (4) exon skipping (21 cases), and (5) composite alternative splicing (different combinations of several alternative splicing events, 11 cases). All cases and the EST evidence are displayed at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/alternativeSplicing/. Intron retention may in part result from inconsequential inefficient splicing or inclusion of incompletely spliced transcripts in EST libraries; thus, evidence for intron retention is not discussed further here (338 cases; listed at (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/alternativeSplicing/intron_retention.html). Based on the EST evidence, we calculated a lower bound for the fraction of alternatively spliced genes as 1.2% (327 out of 27,288). Although EST sampling and coverage remains limited, alternative splicing would seem to be much less pervasive than observed in mammalian systems. For example, if we assume that 5% (or 20%) of the transcripts of an alternatively spliced gene represent the alternative isoform, then the average of seven ESTs per gene result in a 30% (or 79%) detection rate of this gene as alternatively spliced. Thus, limited EST sampling alone should not account for the low estimate of the fraction of alternatively spliced genes.

Mini-exons and mini-introns

Currently there are two non-exclusive models regarding the mechanisms of splicing: intron definition purports interactions of splice site recognition factors across the intron, whereas exon definition suggests interactions of splicing factors at the acceptor and donor sites from consecutive introns across the interspersed exon (Berget 1995). The latter model provides a conceptual framework for the molecular recognition of the very long introns occurring in some mammalian genes, while the former model may be the simplest model for recognition of terminal and short introns. For either model, the existence of very short introns and exons raises difficult questions about the steric accommodation of multiple splicing factors.

Based on EST evidence, we did not find any introns less than 50 bp. According to the GenBank annotation, there are 46 introns ranging from 1-10 bp, but it seems likely that these are annotation mistakes. One 27 bp intron was annotated in the gene At3g53740, which is supported by full-length cDNA CT267357 (gi:21405387). However, 33 pcSPAs uniformly support a continuous exon in that position. It is possible that this region is polymorphic between the Columbia and other ecotypes and that the cognate origin of CT267357 includes a standard size intron.

Conversely, 128 non-terminal mini-exons are supported by EST evidence. These exons range in size from 5-25 bp, with 13 of them no longer than 10 nucleotides in length (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/miniexons.html). In a few cases, these mini-exons may occur in regions of increased alternative splicing activity. An example of this is given in Fig. 5. However, most mini-exons appear to be constitutively spliced, as confirmed by the consistent alignments of several ESTs. For example, a six nucleotide exon in At5g14030 is unanimously confirmed by 12 EST spliced alignments and conserved in an apparent rice homologous gene (Figs. 7 and 8). Due to steric constraint imposed by their size, we find it difficult to explain the accurate splicing of mini-exons by exon-definition, and intron-definition and / or facilitation of splicing by splicing enhancers may be a more plausible splice site selection model in this case. Interestingly, most mini-exons are characterized by high splice prediction scores in the flanking exon-intron junctions (data not shown), suggesting that the associated spliceosome and mechanism of splicing involved in resolving mini-exons may be highly similar to that of normal exons.

DISCUSSION

Expressed Sequence Tags (ESTs) have become the most popular method for gene discovery in eukaryotic species without a whole genome sequencing project as well as a key technology for genome annotation when genome sequence data are available. We are particularly interested in systematic, functional, and phylogenetic comparisons of the gene repertoires of plants. Currently, a near complete genome has been assembled for only Arabidopsis thaliana and rice. In contrast, some of the largest species-specific EST collections are from plants, including wheat (more than 415,000), barley (more than 310,000), soybean (more than 305,000), maize (more than 195,000), and Medicago truncatula (more than 180,000; source: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Kalyanaraman et al. (2003) present a novel algorithm and software program (PaCE) to cluster large sets of ESTs into contigs that represent distinct gene fragments and its application to 22 plant species EST sets. Our motivation for the mapping of Arabidopsis ESTs onto the Arabidopsis genome was in part derived from the need for a confirmed standard of proven EST clusters against which to gauge the success of EST clustering programs that do not incorporate genome sequence data. Here, we derived a number of different standards from uniquely mapped Arabidopsis ESTs depending on the minimal overlap required between different EST spliced alignments. All spliced alignments are displayed at a novel web resource, http://www.plantgdb.org/AtGDB/, which was specifically designed to view and explore all Arabidopsis gene structure annotation and evidence therefore.

In comparison to other indexing methods such as UniGene or the TIGR Gene Indices that work entirely on the mRNA level, genome location based clustering not only has the advantage of accuracy but also allows using low quality ESTs more effectively. For example, EST gi:8332684 has a uhqSPA with a similarity score marginally higher than 0.8, but the GeneSeqer spliced alignment still accurately reveals the exon-intron boundaries of the gene At1g20620 (catalase 3). This EST is clustered with hundreds of other cognate ESTs located in the same region. However, although labeled as weakly similar to At1g20620 it is clustered as a singleton in the TIGR *Arabidopsis* Gene Index.

Surprisingly, the complete EST mapping revealed a large number of discrepancies between the current gene structure annotation and assignments of exons and introns indicated by the spliced alignment. Previously, Haas et al (2002) reported that 1,591 Arabidopsis genes were incorrectly annotated at the time of their comparison with the 5,000 full-length Ceres/TIGR cDNAs, and an additional 240 putative novel genes were identified by the same set of cDNAs. This suggested that full-length cDNA data should greatly improve genome annotation efforts. The most recent release of

Arabidopsis genome annotation from TIGR, used in this study, does incorporate full-length cDNA spliced alignments, thereby reducing the number of contradictory annotations compared to prior annotations. However, there are still about 1,000 genes inaccurately annotated according to our analysis. Furthermore, GeneSeqer alignments using the same full-length cDNA data set as Haas et al. (2002) indicates that for 23 matching genes the current annotation remains erroneous (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/geneAnnotationVScdna.html), suggesting that even with full-length cDNAs, gene identification is still not trivial. Interestingly, about 20% of the gene locations of this representative set of full-length cDNAs are embedded in longer EST alignments (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/extendedCoverage.html). Haas et al. (2002) also reported length differences between the Ceres/TIGR full-length cDNAs (ATcdna set, from ecotypes Wassilewskija and Landsberg erecta) and RIKEN full-length cDNAs (ecotype Columbia; Seki et al. 2002) that may reflect alternative transcription initiation and termination sites, possibly polymorphic among different ecotypes. However, these differences are minor and do not in any other way obscure gene structure prediction. In particular, our comparison with EST spliced alignments show that, except for a few cases, the cDNA confirmed introns are identically predicted by the EST alignments, about 98% of which are with ESTs from the Columbia ecotype. The current sampling of ESTs from different ecotypes is insufficient to assess differences in gene expression or splicing patterns between the ecotypes.

In addition to providing the standards for EST clustering and data for refining basic gene structure annotation, the spliced alignments also provide a rich resource for more in-depth analysis of pre-mRNA processing, including assessment of the extent of alternative splicing and use of non-canonical splices sites. Based on very stringent spliced alignment criteria, we established alternative splicing (excluding possible intron retention) for only about 1.5% of the *Arabidopsis* genes. The majority of alternative splicing occurs at either the donor site or the acceptor site of an intron but not on both ends simultaneously (292 out of 327 cases). We also observe that most alternative splice sites are within 50 bp of the common splice site (220 out of 292). Specifically, in 134 cases, the distances between the alternative splice site and the common splice site is less than 10 bp. Such transcript isoforms with minor difference may easily be overlooked in conventional EST clustering and transcript assembly. For example, the gene At1g02500 has two alternative isoforms with a difference of only three bases in the location of the acceptor site of its sole intron. Each of the isoforms has at least six ESTs to support its unique gene structure. However, all of these ESTs are assembled into one index in the TIGR Arabidopsis Gene Index (Cluster ID:TC149272).

Most certainly, these estimates of the occurrence of alternative splicing are very conservative. First, these estimates were based on only very good spliced alignments that leave no doubt as to the origin of the respective ESTs. Second, the *Arabidopsis* EST collection is still very small compared to the human collection, for example, which is about 30 times larger. However, we still estimate the occurrence of alternative splicing in *Arabidopsis* much lower than the reported 40-60% of human genes (Black 2000; Brett et al. 2002; Modrek and Lee 2002).

Currently, most gene identification efforts rely heavily on *ab initio* gene prediction programs (Pavy et al 1999). However, few *ab initio* gene identification programs successfully make alternative splicing predictions, consider non-canonical splice sites, or other exceptional cases. For example, a special situation where the start codon (ATG) of a gene is interrupted by an intron would confuse almost every *ab initio* gene prediction algorithm currently available. Similarly, mini-exons (EST confirmed examples of which are displayed at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/miniexons.html) will generally be neglected due to their small coding potential, especially if the length of the mini-exon is a multiple of three. Thus it would seem imperative that spliced alignment be a key technology of genome annotation. The GeneSeqer program (Usuka et al. 2000) is very convenient for that purpose.

To facilitate refined genome annotation and further study of pre-mRNA processing based on the spliced alignment data, all of our results were stored in a MySQL database and are visually presented on a special web site, AtGDB (http://www.plantgdb.org/AtGDB/). Several established and comprehensive *Arabidopsis* databases are already available to date, such as TAIR (http://www.arabidopsis.org/), MIPS (http://mips.gsf.de/proj/thal/), and the TIGR *Arabidopsis thaliana* Database (http://www.tigr.org/tdb/e2k1/ath1/). All displays in AtGDB are linked to the corresponding entries in those databases. AtGDB adds a convenient sequence-centered view of the genome. Users of AtGDB can easily find the distribution of target sequences in the genome, see their related annotations, and exact genomic coordinates (based upon the most recent release of *Arabidopsis* genome annotation) of ESTs and cDNAs. Analytical tools are linked to the displays to allow further analysis with additional data, for example spliced alignment with ESTs from sources other than *Arabidopsis*. We hope that this analysis and the new web tools will contribute to more complete and accurate genome annotation.

METHODS

Data Sets

The five chromosome sequences of *Arabidopsis thaliana* were obtained from GenBank (http://www.ncbi.nih.gov/entrez/query.fcgi?db=Nucleotide) as accessions NC_003070 (chromosome I, dated 8-20-2002, 30,028,691 bp), NC_003071 (chromosome II, dated 8-20-2002, 19,646,746 bp), NC_003074 (chromosome III, dated 8-20-2002, 23,467,821 bp), NC_003075 (chromosome IV, dated 8-20-2002, 17,550,036 bp), and NC_003076 (chromosome V, dated 8-20-2002, 26,583,670 bp). *Arabidopsis* ESTs were downloaded from the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/). Our analysis was based on 176,915 EST records available October 25, 2002 (dataset label: ATest). According to the GenBank records, 111,155 non-RIKEN ESTs were derived from the Columbia ecotype. An additional 61,481 ESTs are from RIKEN, and these ESTs were also from Columbia (Seki et al., 2002). Only 337 ESTs are indicated as ecotype Landsberg, and no ecotype information is given for the remaining about 4,000 ESTs. A set of 27,288 putative *Arabidopsis* proteins was obtained from The Institute for Genome Research (TIGR, ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep), which represented the latest annotation of the *Arabidopsis* genome made by TIGR (dataset label: ATpep, version: July 25, 2002). 5,017 full-length cDNAs sequenced by Ceres, Inc. were downloaded from the TIGR ftp site (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/ceres/Ceres.arab.cdna). Only the subset of 5,000 sequences deposited in GenBank (Entrez search: Arabidopsis [ORGN] AND FLI_CDNA [KYWD] AND Haas [AUTH]) were used in this study (dataset label: ATcdna, version: March 2, 2001). These cDNAs were derived from the Wassilewskija and Landsberg *erecta* ecotypes (Haas et al, 2002).

EST Mapping by Spliced Alignment

Alignment of cDNAs or ESTs to a genomic template is known as spliced alignment, because the alignment must correctly reflect the removal of introns from the pre-mRNA copy of the genomic template. Several programs and services are available for this task, including PROCRUSTES (Gelfand et al, 1996), NAP (Huang et al, 1997), SIM4 (Florea et al, 1998), est_genome (Mott, 1997), Spidey (Wheelan et a., 2001), and GeneSeqer (Usuka et al, 2000; Usuka and Brendel, 2000). The alignments discussed here were derived with the GeneSeqer program. The program involves pre-processing of the cDNA/EST set to generate a suffix array of these sequences, subsequent fast

matching of cDNAs/ESTs to the genome based on significant blocks of sequence identity, and spliced alignment by dynamic programming based on predicted splice site probabilities and sequence similarity scores. Using default parameter settings, the entire mapping of ATest was achieved in about 120 hours on a 1-GHz Pentium Pro III processor CPU.

Selection of High-quality EST Alignments

The default GeneSeqer parameters are set to allow detection of gene structure through alignment of ESTs from non-cognate ESTs derived from a homologous gene elsewhere in the genome (or even ESTs from a homologous locus in a related species). For some of the questions studied here, it was necessary to restrict the data to only the cognate alignments. Because of allelic variation and sequencing errors, even cognate alignments will not necessarily display 100% sequence matching, however the overall alignment quality generally should be much higher than for heterologous alignments. For a given EST, GeneSeqer assesses alignment quality by two parameters: a similarity score, defined as the ratio of the observed alignment score over the maximum possible alignment score obtained in the absence of any substitutions and insertions or deletions; and a coverage score, defined as the fraction of the EST nucleotides involved in the displayed alignment (because the GeneSeqer spliced alignment is local, any poorly matching N- or C-terminal EST regions are culled from the displayed alignment) . Here, we define high-quality EST spliced alignments (**hqSPA**) as alignments that give similarity and coverage scores both at least 0.8. ESTs with at least one hqSPA are defined as **hqEST**. A hqEST is further categorized according to the number of hqSPAs derived from the given EST. It is called a unique hqEST (**uhqEST**) if the EST matches a unique locus in the genome, and it is called a **mhqEST** if the EST matches multiple sites in the genome (presumably corresponding to duplicated genes). The corresponding spliced alignments are referred to as **uhqSPAs** and **mhqSPAs**.

The major task of spliced alignment discussed in this paper was to identify cognate positions for each entry of ATest. Because the EST set was not masked or filtered to remove contaminations, low complexity regions or repeats, and because high sensitivity / low specificity default GeneSeqer parameters were applied for the spliced alignment, we limited most of our derived results to hqSPAs and hqESTs. The product of similarity and coverage scores was utilized as a measure to identify the putative cognate alignments (**pcSPA**), based on the assumption that the pcSPA should have the best score among hqSPAs for each specific hqEST. Due to recent gene duplications, possible genome

assembly errors, or other uncertain reasons, some hqESTs may have several hqSPAs with identical or near-identical score in different locations of the genome. Thus, the pcSPA for each hqEST is not necessarily unique. The distribution of score differences among multiple hqSPAs for an EST is shown in Fig. 3. Based on this distribution, all hqSPAs with score strictly within 0.015 of the maximal score for that EST were labeled as pcSPA. With default parameters, a GeneSeqer reported similarity score of s corresponds to $0.5 \cdot (1+s) \cdot 100\%$ sequence identity (for an alignment without gaps). Thus, two alternative full-length alignments of an EST will be distinguished as cognate and non-cognate if the weaker match has on average one additional mismatch to the genomic sequence per 100 nucleotides compared to the better match. The average nucleotide difference between the duplicated genes identified by hqESTs was calculated as $11.4 \pm 4.6\%$. Therefore the given criterion would safely distinguish duplicated genes except for very recent duplications that result in such minor sequence differences that they are indistinguishable from EST sequencing error rates.

EST Clustering and Assembly

hqESTs were mapped to the *Arabidopsis* genome based on pcSPAs as described in the previous section. The mapped hqESTs were clustered according to genome coordinates derived from their pcSPAs requiring a defined minimal overlap length or a maximal coverage gap size. Precisely, let est1 map to region [a,b] and est2 to region [c,d], where a<=c, on the same chromosome; then est1 and est2 are clustered if c<=b+G+1, where G is the clustering parameter. G could be negative (overlap required) or positive (specifying the maximal allowed gap). For ESTs giving multiple exon spliced alignments, the overlap rule is superceded by the requirement for consistency of strand orientation as indicated by GeneSeqer. Thus, ESTs from overlapping genes in opposite transcriptional directions can be separated into different clusters (*cf.* Fig. 5). Additionally, ESTs from the same plasmid (clone pairs) were used to join clusters independent of their local map coordinates. Different sets of clusters based on alignment and clustering parameters are available at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/ESTclustering.html. ESTs of each cluster were further assembled by the built-in function of GeneSeqer to generate alternative gene structures (AGSs) and predicted peptide sequences (PPSs) derived from long open reading frames in the AGSs. The PPSs were searched against ATpep via BLASTP to locate putative novel genes as described in Analysis of EST Spliced Alignments below.

Quality Control

The set of full-length cDNAs was aligned to the genome similarly to the EST alignments (the GeneSeqer option –x 30 –y 50 was used which probes for potential gene locations by about 50-base identities in the suffix array, thus quickly identifying cognate loci). These alignments served as quality control in two ways. First, the results test the integrity of our analysis method. Because these full-length cDNAs were previously used to improve the *Arabidopsis* genome annotation (Haas et al. 2002), the cDNA spliced alignments from GeneSeqer are expected to be consistent with the genome annotation. Second, we can check whether the pcSPAs are consistent with the cDNA alignments in regions of overlap. The 5,017 full-length cDNAs can be regarded as a random sample of the total gene set of *Arabidopsis*. Comparing the coverage of ESTs relative to these cDNAs tests the limits of EST projects.

Database and Web Interface

The raw output of GeneSeqer occupied a total of 1.6 billion bytes disk space. The output was parsed and imported into a MySQL relational database management system (http://www.mysql.com) for further analysis. The database is accessible via the web at http://www.plantgdb.org/AtGDB/. Supplementary data for the results of this study are available at http://www.plantgdb.org/AtGDB/prj/ZSB03PP/.

High-quality Predicted Introns

GeneSeqer gives two scores to each splice site, a prediction score and a local similarity score. The prediction score is between 0 and 1.0, based on a statistical model for the probability of the site to function as a splice site. Non-canonical splice sites receive 0 as prediction score. The local similarity score measures sequence matching in the 40-50 bp flanking exon regions derived from the spliced alignment. This score is also normalized to 1.0 for complete identity. For exons shorter than 40 bp, the local similarity scores of the flanking splice sites are both set to 0. In this study, high-quality predicted introns were selected as predicted introns with (a) splice site prediction scores for the donor and acceptor sites both higher than 0, i.e., the intron should be a canonical intron; and (b) local similarity scores for the donor and acceptor sites both higher than 0.95 (implying, that the flanking

exons should be no less than 40 bp and at most one mismatch is allowed in the 40-50 bp flanking exon region alignment).

Identification of U12-introns

The 5'-site motif ATCC in positions +3 to +6 is highly conserved in U12-introns (Wu and Krainer, 1996; Sharp and Burge, 1997; Burge et al., 1998), where the numbering +1 to +6 denotes the first six nucleotides of the intron starting at the 5'-splice site. On the basis of this observation, we selected one AT-AA and 153 GT-AG introns as potential U12-class introns among all the EST-confirmed introns (in addition to the 23 U12-dependent AT-AC introns discussed in the text). To further classify these sites, we used a procedure similar to those described by Burge et al. (1998) and Levine & Durbin (2001). First, MEME (Bailey & Elkan, 1994) was used to define motifs for the donor and branch sites of the 23 manually verified U12-class AT-AC introns. These motifs were then used to query the additional 154 candidates via the MAST application (E-value threshold set to 1.0; Bailey & Gribskov, 1998) and 18 introns with motif E-values less than 1.0 for both motifs were characterized as likely U12-introns.

Analysis of EST Spliced Alignments

The mapped ESTs provide a rich data set for studying many aspects of genome and gene structure. Here, we have explored the following issues. 1) Consistency of gene structure annotation. EST spliced alignments reveal partial or full gene structures and are thus helpful to check and refine *ab initio* gene predictions (Brendel and Zhu 2002). Distinct introns derived from all EST alignments were utilized to identify what fraction of annotated introns is supported by EST evidence. Only high-quality predicted introns were used to identify annotated introns that are not supported but contradicted by EST evidence. Even in the well-annotated *Arabidopsis* genome, there may still be some genes that are not yet described. We defined putative (partial) novel genes as EST-derived conceptual transcripts with an open reading frame longer than 300 bp (PPSs with more than 100 amino acid residues), but displaying no significant similarity to proteins in ATpep (threshold 1e-10 using BLASTP) and have no overlap with annotated genes. 2) 5'- and 3'- Untranslated Regions (UTRs) in mRNAs. We used the EST evidence to identify UTR exons and introns. 3) Non-canonical splice sites. Non-canonical splice sites obtain a prediction score of 0 in the GeneSeqer spliced

alignments. Therefore it is very simple to identify potential non-canonical introns. To exclude questionable spliced alignments and remove redundancy, only distinct introns with flanking exons of at least 40 bp and local similarity score greater than 0.95 were selected for further analysis and categorization according to the observed intron borders. 4) Alternative splicing. All high-quality introns were mutually compared to find overlapped but non-identical introns, indicating different types of alternative splicing (except intron retention, cases of which were identified separately). 5) Mini-exons and mini-introns. Mini-exons were selected from hqSPAs containing at least one exon of at most 25 bp, with 100% alignment identity over the entire exon region and canonical splice sites as boundaries. Similar criteria were also applied to seek mini-introns not exceeding 50 bp in length.

ACKNOWLEDGEMENTS

The authors would like to thank P. Vedell for early contributions to this work and J. Schlueter for critical reading of the manuscript. This work was supported in part by NSF grant DBI-0110254 to V.B.

LITERATURE CITED

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2:** 28-36

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics **14:** 48-54

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem **270:** 2411-2414

Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-

genome biology. Cell **103:** 367-370

Bouck J, Yu W, Gibbs R, Worley K (1999) Comparison of gene indexing databases. Trends Genet **15:** 159-162

Brendel V, Kleffe J (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucleic Acids Res **26:** 4748-4757

Brendel V, Zhu W (2002) Computational modeling of gene structure in *Arabidopsis thaliana*. Plant Mol Biol **48:** 49-58

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. Nat Genet **30:** 29-30

Brown JW, Smith P, Simpson CG (1996) *Arabidopsis* consensus intron sequences. Plant Mol Biol **32:** 531-535

Burge CB, Padgett RA, Sharp PA (1998) Evolutionary fates and origins of U12-type introns. Mol Cell **2:** 773-785

Burset M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res **28:** 4364-4375

Burset M, Seledtsov IA, Solovyev VV (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res **29:** 255-259

Coward E, Haas SA, Vingron M (2002) SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. Trends Genet **18:** 53-55

Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. Nat Genet **29:** 412-417

Davuluri RV, Suzuki Y, Sugano S, Zhang MQ (2000) CART classification of human 5' UTR sequences. Genome Res **10:** 1807-1816

Dietrich RC, Incorvaia R, Padgett RA (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol Cell **1:** 151-160

Fernandes J, Brendel V, Gai X, Lal S, Chandler VL, Elumalai RP, Galbraith DW, Pierson EA, Walbot V (2002) Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization. Plant Physiol **128:** 896-910

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res **8:** 967-974

Gelfand MS, Mironov AA, Pevzner PA (1996) Gene recognition via spliced sequence alignment. Proc Natl Acad Sci **93:** 9061-9066

Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg, SL (2002) Full-length messenger RNA sequences greatly improve genome annotation. Genome Biology **3:** research0029.1-0029.12

Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. Genomics **46:** 37-45

Huang YH, Chen YT, Lai JJ, Yang ST, Yang UC (2002) PALS db: Putative Alternative Splicing database. Nucleic Acids Res **30:** 186-190

Kalyanaraman, A, Kothari S, Brendel V, Aluru, S (2003) Efficient clustering of large EST data sets on parallel computers. Nucleic Acids Res **31:** in press

Kan Z, Rouchka EC, Gish WR, States DJ (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res **11:** 889-900

Levine A, Durbin R (2001) A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res 29: 4006-4013

Modrek B, Lee C (2002) A genomic view of alternative splicing. Nat Genet **30:** 13-19

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci **13:** 477-478

Pavy N, Rombauts S, Déhais P, Mathé C, Ramana DVV, Leroy P, Rouzé P (1999) Bioinformatics **15:** 887-899

Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. Nucleic Acids Res **30:** 335-340

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res **29:** 159-164

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y et al. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science **296:** 141-145

Sharp PA, Burge CB (1997) Classification of introns: U2-type or U12-type. Cell **91:** 875-879

Tabaska JE, Davuluri RV, Zhang MQ (2001) Identifying the 3'-terminal exon in human DNA. Bioinformatics **17:** 602-607

Usuka J, Brendel V (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J Mol Biol **297:** 1075-1085

Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic

DNA template. Bioinformatics **16:** 203-211

Wheelan SJ, Church DM, Ostell, JM (2001) Spidey: a tool for mRNA-to-genomic alignments. Genome Res **11:** 1952-1957

Wu HJ, Gaubier-Comella P, Delseny M, Grellet F, Van Montagu M, Rouzé R (1996) Non-canonical introns are at least 10(9) years old. Nat Genet **14:** 383-384

Wu Q, Krainer AR (1996) U1-mediated exon definition interactions between AT-AC and GT-AG introns. Science **274:** 1005-1008

Wu Q, Krainer AR (1999) AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol Cell Biol **19:** 3225-3236

Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. Genome Res **11:** 803-816

Figure Legends

Figure 1. Classification of *Arabidopsis* ESTs based on spliced alignment quality. Of a total of 176,915 ESTs, 2,059 ESTs have no significant hits in the *Arabidopsis* genome, 4,968 ESTs have only low-quality spliced alignments (lqEST), and the remaining 169,888 ESTs have high-quality spliced alignments. The latter category consists of 146,527 ESTs that match a unique (i.e., their cognate) locus in the genome (uhqEST) and 23,361 ESTs that have multiple high-quality spliced alignments (mhqEST), representing different loci of duplicated genes or multigene families.

Figure 2. Distribution of the 170 hqSPAs for EST gi:9787698 on the *Arabidopsis* genome. Each chromosome is represented by two dark green bars, with the centromere marked by a space between the horizontal bars. Locations of the spliced alignments are shown by red bricks. Almost all hits are around the centromeres. Alignment scores suggest that the EST originates from the 12,075,567-12,075,806 bp region on chromosome three (marked by the green arrow). This EST shows high similarity with *Arabidopsis* gene At1g38360 (gi:18426880), a putative retroelement polyprotein gene. This display is shown as an example of the visualization tools at AtGDB which will dynamically generate similar graphics for any set of GenBank gi accessions or genes matched by common descriptions.

Figure 3. Distribution of the score differences between maximal and submaximal scoring hqSPAs for mhqESTs. Each hqSPA is scored by the product of similarity and coverage values (see METHODS). Most of the score differences fall in the range 0.08-0.20. Based on the displayed distribution, a critical value 0.015 was set such that each hqSPA with score difference smaller than 0.015 compared to the maximal scoring hqSPA for a given EST is designated as putative cognate spliced alignment (pcSPA), representing the likely origin of this specific EST in the genome.

Figure 4. Histogram showing the distribution of pcSPA similarity, coverage, and combined scores.

Figure 5. Visual assessment of EST clustering and gene characteristics for a region of the *Arabidopsis* genome. In the display, which is available for all genomic regions at http://www.plantgdb.org/AtGDB/, pcSPAs originating from EST spliced alignments are shown in red and non-pcSPAs in pink. For multi-exon alignments, the arrow indicates the direction of transcription, inferred from the implied splice site patterns (Usuka et al., 2000). Multi-exon 5'-ESTs are marked by green color at their 5'-terminus, and multi-exon 3'-ESTs are marked by blue color at their 3'-terminus. Single exon ESTs have corresponding 5' / 3' labels at the center of their representations. Pairs of 5'- and 3'-ESTs from the same clone are grouped by green boxes. PcSPAs originating from cDNA spliced alignments are shown in light blue and non-pcSPAs in grey. Dark blue gene structures represent the current GenBank gene annotations for this region. The 5'- and 3'-boundaries of the corresponding coding regions are indicated by green and red triangles, respectively. Note that the current annotation misses the gene represented by clone pair ESTs gi:19867004 & gi:19822861 and gi:19878951 & gi:19799838. The purple structures represent the spliced alignments of TIGR Arabidopsis Gene Index tentative contigs. The figure also shows an alternatively spliced internal mini-exon. This exon of 16 nucleotides occurs in the 5'-UTR of At4g38510, an H+-transporting ATPase (EC 3.6.1.35). The transcript isoform including this intron is supported by ESTs gi:9785303 and gi:8722457. In the same region, EST gi:9787070 supports a different internal exon of 73 nucleotides, and EST gi:19867985 (equal to RAFL-15010615) indicates an alternative transcription start. Note that all sequence records at AtGDB are identified by their unique GenBank gi identifiers. The Riken Arabidopsis Full-Length (RAFL) cDNAs (Seki et al. 2002) indicated thus as RAFL-15451093, RAFL-18377451, RAFL-20268790, RAFL-21689814, RAFL-15010783, RAFL-14517367, RAFL-16323357, RAFL-15010615, and RAFL-19699257 correspond to clones

RAFL05-11-M12, U16016, RAFL06-81-F18, U11966, RAFL03-01-G10, RAFL04-09-A19, U12748, RAFL07-17-H08, and U12937, respectively.

Figure 6. Spliced alignment of *Arabidopsis* EST gi: 5839990 with **A)** the *Arabidopsis* At3g53520 gene encoding a dTDP-glucose 4-6-dehydratase like protein, and **B)** a rice genomic sequence (accession number: AP003271). The two alignments reveal conserved gene structure between *Arabidopsis* and rice, including a conserved AT-AC intron. **C)** Pairwise alignment of the orthologous AT-AC intron sequences. The conserved donor site (ATATCCTY) and branch site motifs (TCCTTRAY) are highlighted in red color.

Figure 7. Visualization of an annotated, normally expressed internal mini-exon. The exon of six nucleotides found in the 3'-coding region of At5g14030 (encoding an unknown protein) is supported by 12 different EST spliced alignments. Strikingly, this miniature exon is also conserved in what appears to be a rice homolog of this gene (see Fig. 8). Symbols are as in Fig. 5. The three cDNAs identified by GenBank gi as CT-21404330, RAFL-14517445, RAFL-22136543 correspond to Ceres/TIGR full-length cDNA 16313 and RAFL clones RAFL02-05-J08 and U12778, respectively.

Figure 8. Evolutionary conservation of a mini-exon. **A)** Spliced alignment of the translated open reading frame (bottom lines) originating from the EST cluster shown in Fig. 7 with a rice genomic clone (GenBank accession number: AP003727); the alignment was made with the GeneSeqer program (Usuka and Brendel 2000). **B)** Alignment of the *Arabidopsis* mini-exon and its flanking introns with a homologous region of the rice genome. The mini-exon is highlighted in red characters, the intron donor sites in green, and the intron acceptor sites in blue.

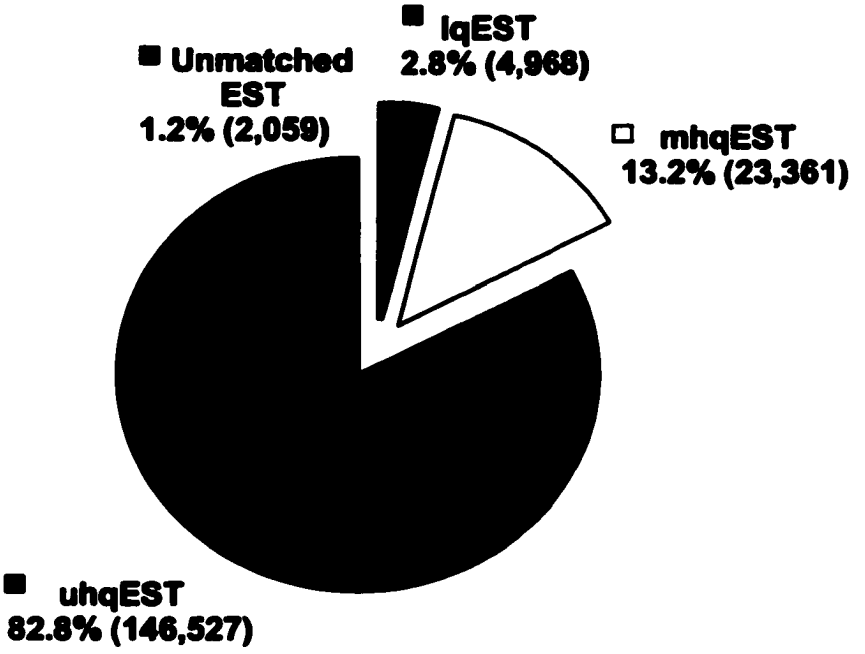

lqEST
2.8% (4,968)
Unmatched EST
1.2% (2,059)
mhqEST
13.2% (23,361)
uhqEST
82.8% (146,527)

Figure 1

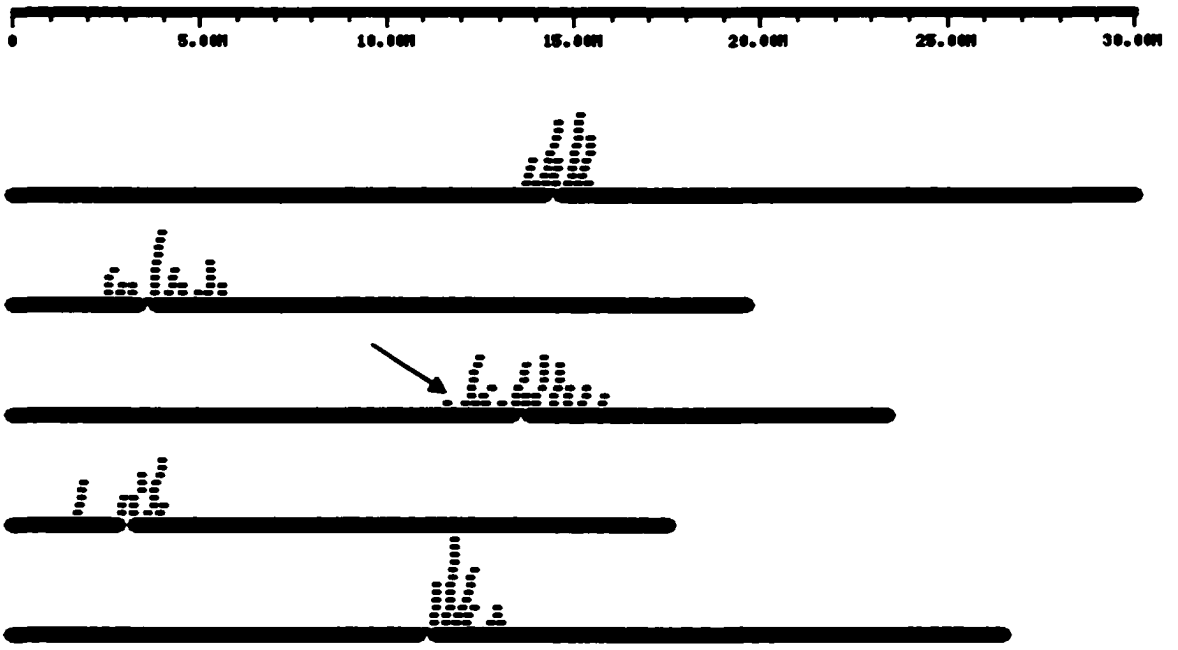


Figure 2

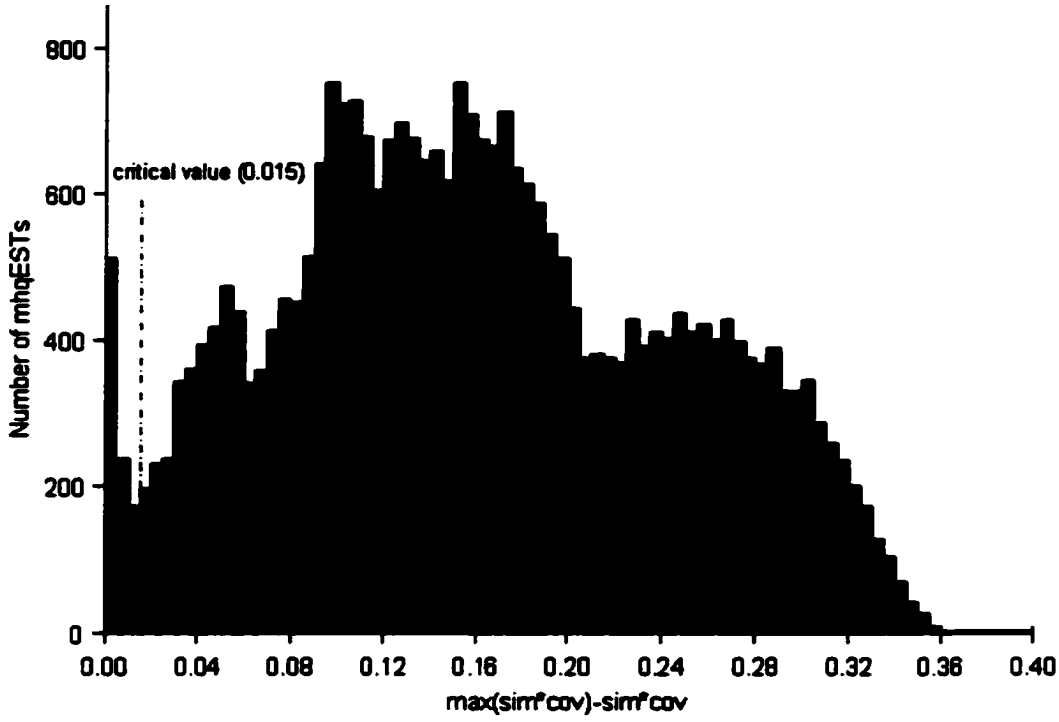

800
600
400
200
0
Number of mhqESTs
critical value (0.015)
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
0.36
0.40
max(sim*cov)-sim*cov

Figure 3

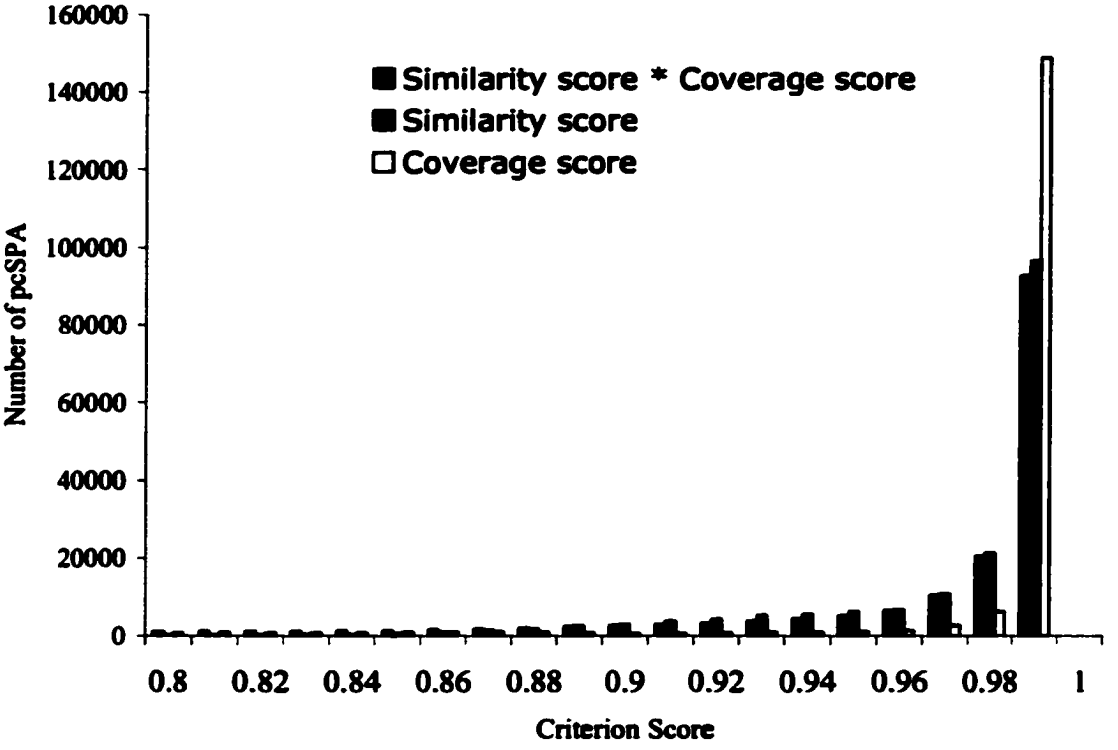

160000
140000
120000
100000
80000
60000
40000
20000
0
Number of pcSPA
Similarity score * Coverage score
Similarity score
Coverage score
0.8
0.82
0.84
0.86
0.88
0.9
0.92
0.94
0.96
0.98
1
Criterion Score

Figure 4

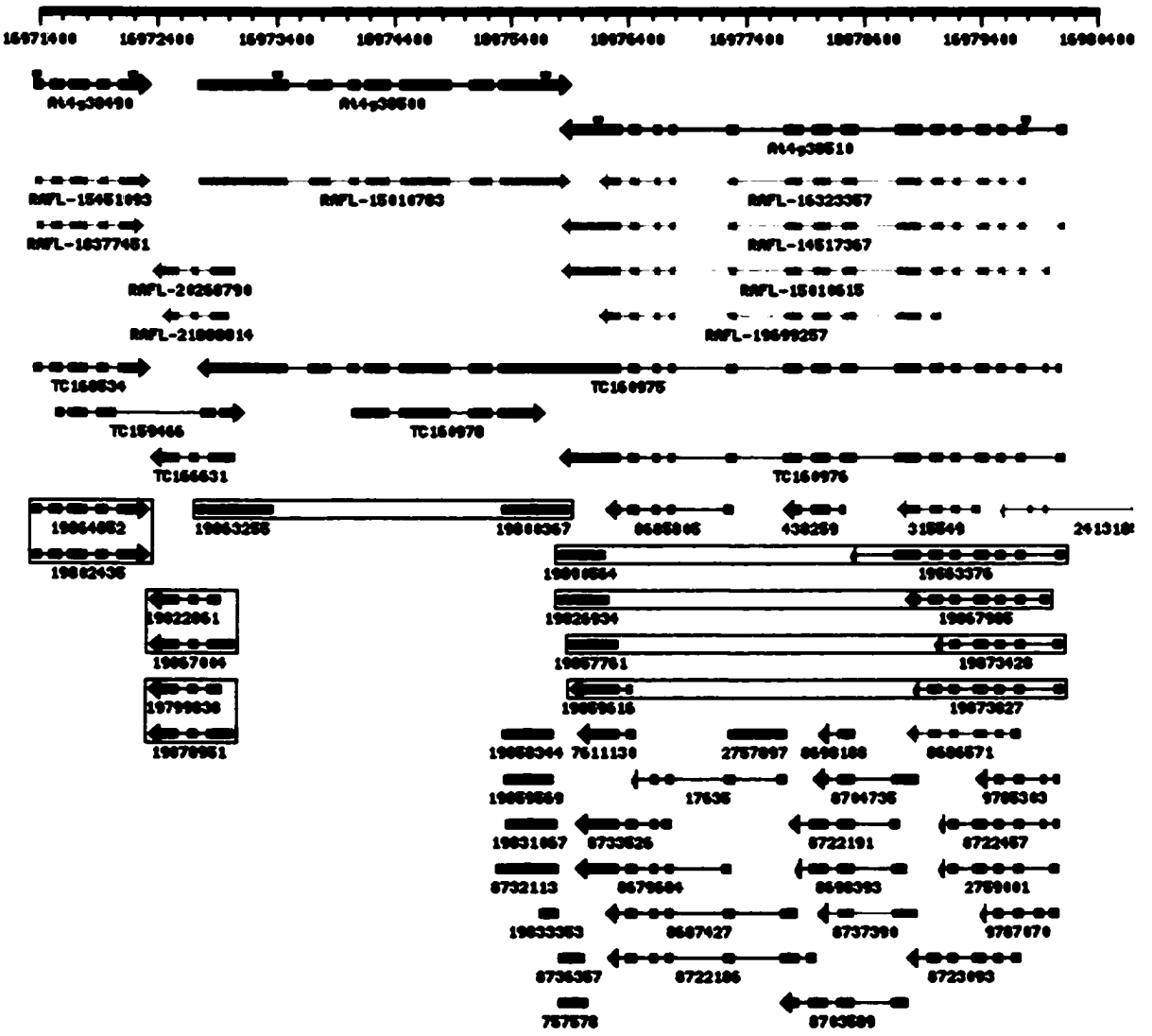


Figure 5

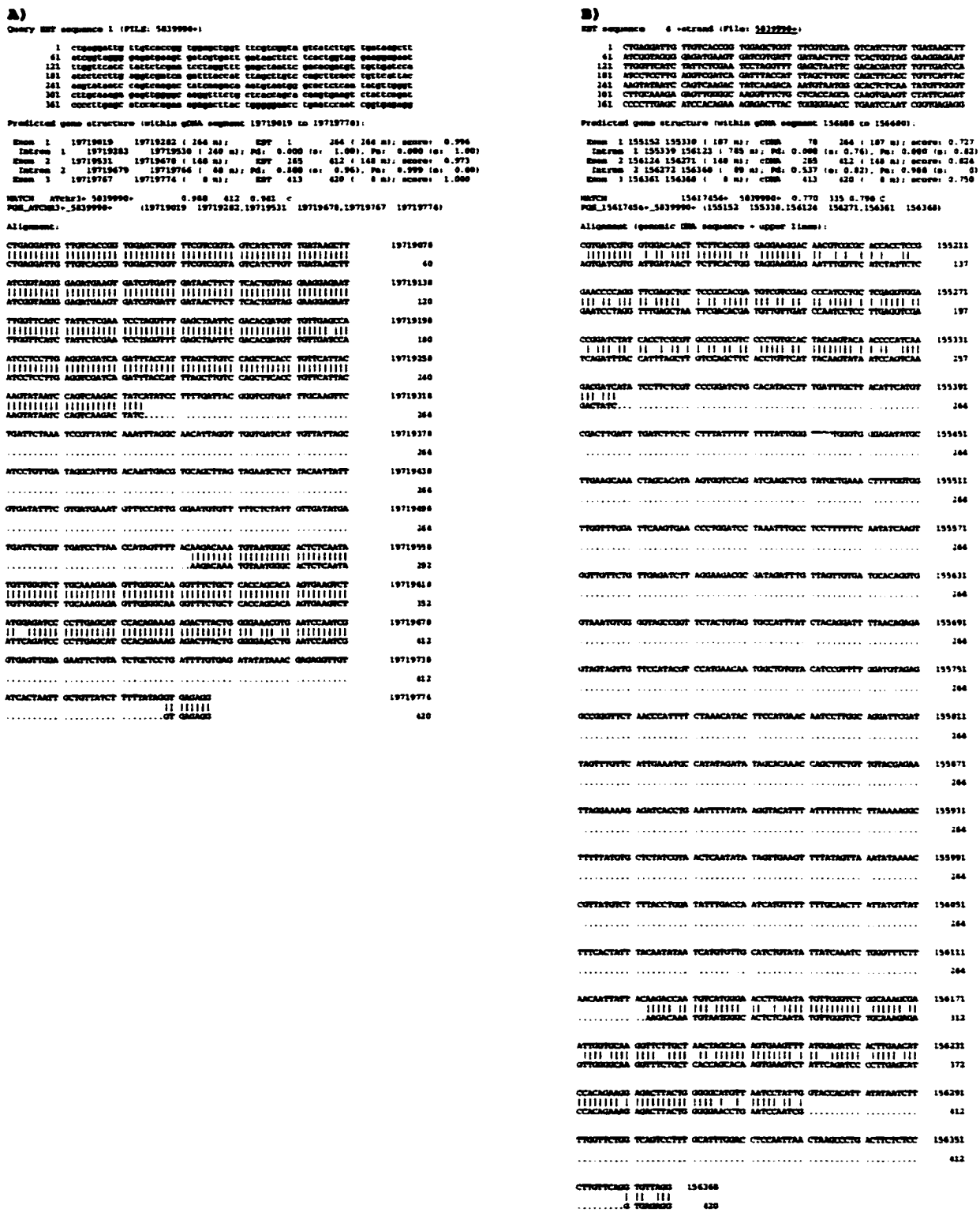


Figure 6

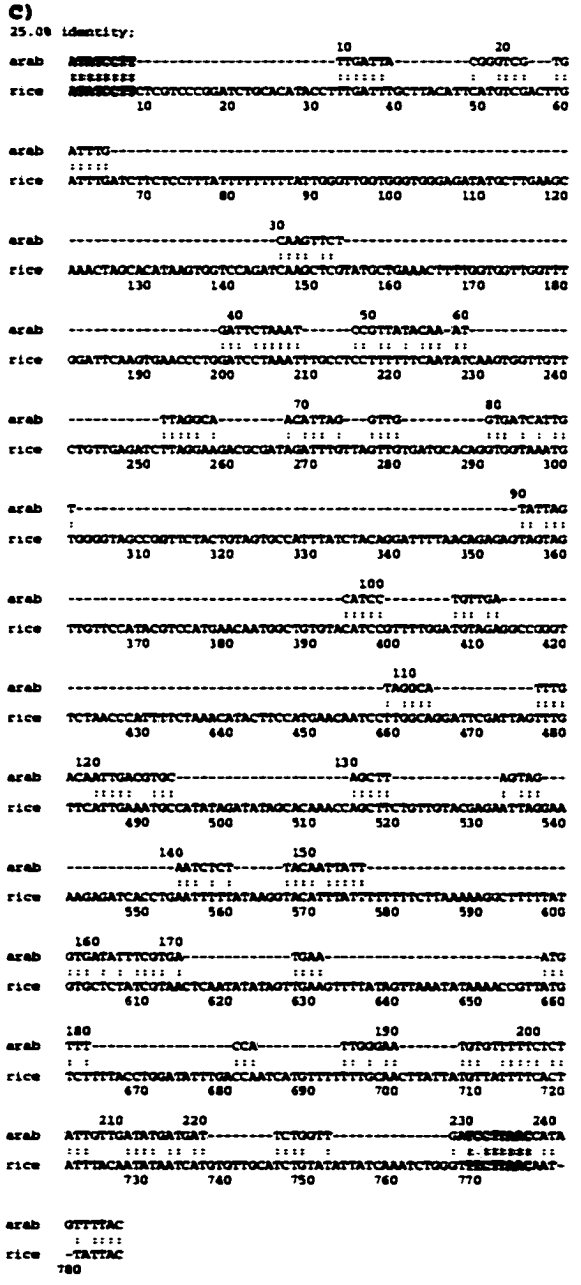


Figure 6 (Cont.)

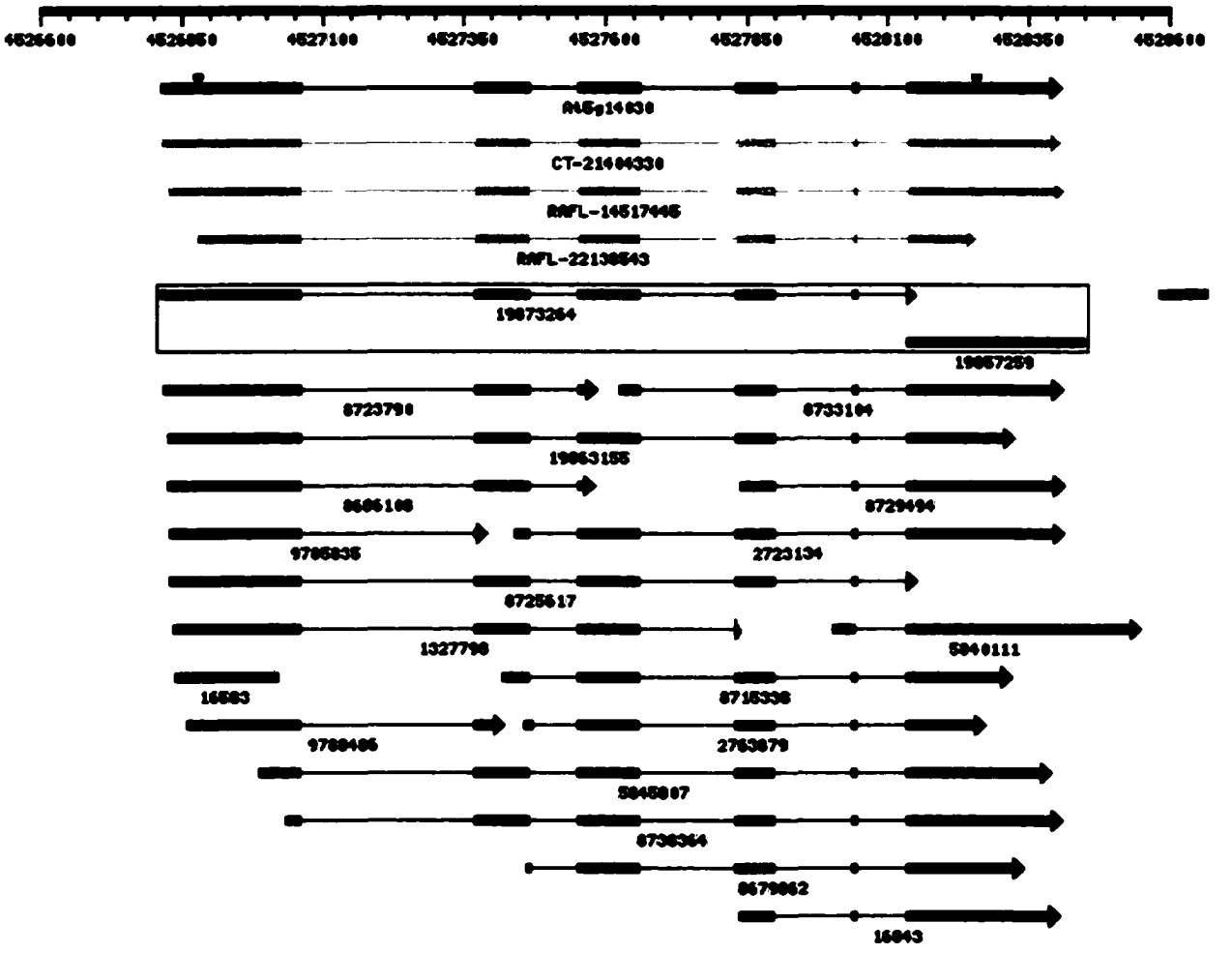


Figure 7

A)

```
CCATCCTTGC ACTTGATGTT CTTGCTGAGA GGCCTCCAGA AAAGAAGTTT GAATGGGTAA   146733
P  I  L  A   L  D  V   L  A  E   R  P  P  E   K  K  F   E  W
|  *  |      |  |  *   |  |  *   *  |  |      |     .   *
P  L  L  P   L  D  I   L  A  D   K  P  P  T   K  P  L   D  V ....     174

GTAATTGTGC CATTGCATGT ACATTCTAGT CTGTTTCGCT CTATGATCCT AATGTGCTGA   146793

.......... .......... .......... .......... .......... ..........     174

TGCTGTTTAT GTTTCTGGCA TGCGCCGTGT ATGAATTCTT GATGTTCTCT CCTTGTTGGT   146853

.......... .......... .......... .......... .......... ..........     174

TCAGGCTAAG GTAAATTTAC TCTTTGTTTT ATTCTACAAG GAGCAACTTT TCTAGTGCAT   146913
     A  K
     |  |
.... A  K  .......... .......... .......... .......... ..........     176

CTATGGTAGA AGAAATTACT GAGAAATTAT CCATTTGCAG AGGCTTGTGG CGAAGTACGG   146973
                                           R  L  V   A  K  Y  G
                                           |  |  *   |  |  |  |
.......... .......... .......... .......... R  L  L   A  K  Y  G     183

GTCGCTGGTG TCCGTTGTTG GCTTG   146998
  S  L  V   S  V  V   G  L
  |  |  |   |  |  *   .  *
  S  L  V   S  V  I   S  M       191
```

B)

```
                10        20        30        40        50        60
arab_f GTAAATTTGACCTTCTAAGGACTATTCTTTCAATCATCTAGTCAGAGTTTTTTTCTTTTC
       :::: :       :: :      : :::    :   :   :::::::    ::::    ::: :
rice_f GTAAGTAA----TTGTG----CCATTGCATGTACATTCTAGTCT--GTTTCGCTCTATGA
                    10            20         30           40          50

              70          80         90          100       110
arab_f TTCTCAATAT--TGACTTTCATTGCTATCTTCTGT--TGTGTCTTGGATGTTGCTCTTGA
       : :: ::: :  ::: :     ::  : : :::::   :: : : :: :::     ::::::
rice_f TCCT-AATGTGCTGA-TGCTGTTTATGT-TTCTGGCATGCGCCGTGTATGAAT-TCTTGA
                60          70          80         90        100

        120       130         140       150         160       170
arab_f TGCTGCCAAAATAATC--TCAGGCCAAGGTAACTCAATGAC-CGATGTGTAATTCTCACT
       :: :  :     :  :   :::::: :::::::    : : :: :  ::: : ::::: ::
rice_f TGTTCTCTCCTTGTTGGTTCAGGCTAAGGTAA---ATTTACTCTTTGTTTTATTCT-ACA
        110       120       130          140       150         160

            180       190         200       210          220
arab_f TCCCAGTGAAATCATCAAAT--ATATATCAGTTTCTCATATTACCGTG----TTTCAATT
          ::  :: :  :: : :  :: ::: ::      : ::::: : :    : :: ::
rice_f AGG-AGC-AACTTTTCTAGTGCATCTATG-GTAGAAGAAATTACTGAGAAATTATCCAT-
              170       180         190       200       210

       230
arab_f GTTGCAG
        ::::::
rice_f -TTGCAG
       220
```

Figure 8

Table 1. Distribution of the number of hqSPAs per hqEST.

# hqSPAs	# ESTs
1	146,527
2	16,116
3	3,697
4	2,235
5	945
6	196
7	68
8	46
9	20
10-19	22
20-99	14
164	1
170	1
Total	169,888

All hqESTs (Fig. 1) were classified according to their number of hqSPAs. The chromosomal distribution of the 170 hqSPAs of EST gi:9787698 is displayed in Fig. 2.

Table 2. Effect of clustering criterion on the number of EST clusters.

Clustering Criterion	# Clusters
≥ 50 bp overlap	30,154
≥ 0 bp overlap	28,883
≤ 50 bp gap	27,787
≤ 60 bp gap	27,611
≤ 100 bp gap	26,956

ESTs were clustered based on their genomic locations, derived from pcSPAs. Several clustering parameters were tested, ranging from requiring a minimum of 50 bp overlap between clustered ESTs to a maximal gap of 100 bp between clustered EST ends.

Table 3. Distribution of EST cluster size.

Cluster Size	# Clusters
1	9,488
2	4,977
3-4	4,927
5-8	3,971
9-16	2,378
17-32	1,132
33-64	472
65-128	185
129-256	60
257-512	113
513-1024	8
Total	27,611

Cluster size is given in number of ESTs. The displayed numbers are based on the clusters derived with the criterion of 60 bp maximal gap (Table 2).

Table 4. Non-canonical introns (NN represents any dinucleotide)

Type	Number
GC-AG	453
NN-AG (not including GC-AG and GT-AG)	99
GT-NN (not including GT-AG)	80
AT-AC	25
GC-NN (not including GC-AG)	14
Others (26 patterns, each with less than 6 hits)	67
Total	738

The intron types were assigned by the terminal intron dinucleotides based on high-quality spliced alignments.

CHAPTER 6. IDENTIFICATION, CHARACTERIZATION AND MOLECULAR PHYLOGENY OF U12-DEPENDENT INTRONS IN THE *Arabidopsis thaliana* GENOME

A paper has been submitted to *Nucleic Acid Research*[1]

Wei Zhu[2] and Volker Brendel[3]

ABSTRACT

U12-dependent introns are spliced by the minor U12-type spliceosome and occur in a variety of eukaryotic organisms, including *Arabidopsis*. In this study, a set of putative U12-dependent introns was compiled from a large collection of cDNA/EST-confirmed introns in the *Arabidopsis thaliana* genome by means of high-throughput bioinformatic analysis combined with manual scrutiny. A total of 158 distinct U12-type introns were identified based upon stringent criteria, many more than the total number of U12-type introns previously reported for plants. Of particular note is the discovery that the distance between the branch site adenosine and the acceptor site ranges from 10 nt to 39 nt, significantly longer than the previously postulated limit of 21 bp. Further analysis indicates that, in addition to the spacing constraint, the sequence context of the potential acceptor site may have an important role in the 3′ splice site selection. Several alternative splicing events involving U12-type introns were also captured in this study, providing direct evidence that U12-dependent acceptor sites can also be recognized by the U2-type spliceosome. Furthermore, phylogenic analysis accurately dated the fusion event of the two spliceosomes as occurring one billion years ago, subsequent to the divergence of AtNHX1-4 (*Arabidopsis* Na+/H+ antiporter 1-4) and AtNHX5-6.

---

[1] Submitted on Mar. 27, 2003

[2] Primary researcher and author for correspondence, graduate student, Department of Zoology and Genetics, Iowa State University.

[3] Author, Professor, Department of Zoology and Genetics, Department of Statistics, Iowa State University

INTRODUCTION

U12-dependent introns, initially discovered by (Jackson 1991) and (Hall and Padgett 1994), are a class of low-abundance introns which are spliced by the minor class (U12-dependent) spliceosome and are distributed in vertebrates, insects and plants (Burge et al. 1998). This rare class of introns is characterized by highly conserved consensus sequences at the donor and branch sites (Hall and Padgett 1994), in contrast to the much degenerate splice signals in the major class (U2-type) introns that are spliced by the U2-type spliceosome (recently reviewed by Burge et al., 1999; also see Krainer and Wu, 1999). Correspondingly, the U12-type spliceosome consists of specific U4atac, U6atac, U11 and U12 snRNAs that recognize the U12-type splice signals (Hall and Padgett 1996; Tarn and Steitz 1996a; Tarn and Steitz 1996b). In addition, U12-type introns lack a polypyrimidine tract between the branch point sequence (BPS) and the 3' splice site (3'ss). Despite these differences, the U12-type spliceosome resembles the conventional spliceosome in many ways (reviewed by Burge et al. 1999). For instance, irrespective of the lack of sequence similarity, U11, U12, U4atac and U6atac snRNAs are likely to have roles in the U12-dependent spliceosome that are analogous to the roles of U1, U2, U4 and U6 snRNAs in the U2-dependent spliceosome, respectively. Recent experimental data proved that the stem-loop structure within the U6 snRNA can functionally substitute the U6atac snRNA stem-loop (Shukla and Padgett 2001). Moreover, not only is U5 snRNA common in each of the two spliceosomes, a growing number of proteins have been confirmed to be shared by both spliceosomes (Will et al. 1999; Luo et al. 1999; Will et al. 2001; Schneider et al. 2002). U12-type introns typically coexist with U2-type introns in alternate patterns in the same gene (Burge et al. 1998; Levine and Durbin 2001), and the splicing efficiency of U12-type introns can be promoted by splicing the flanking U2-type introns via the exon definition mechanism (Wu and Krainer 1996; Hastings and Krainer 2001; Dietrich et al. 2001b). U11/U12 di-snRNAs were found to bridge the 5' splice site and the BPS in the initial recognition of U12-type introns, suggesting that the mechanism of intron-definition also functions in the splicing of minor introns (Frilander and Steitz 1999).

The U12-dependent spliceosome might have coexisted with the conventional spliceosome in the common ancestor of higher eukaryotes (Wu et al. 1996). The fact that vertebrates and higher plants share conserved features in the functional regions of U6atac and U12 snRNAs also provides evidence indicating an early origin (Shukla and Padgett 1999). The difference of the two splicing machineries implies that the two spliceosomes evolved parallel to each other in separate lineages and then

merged together prior to the divergence of the animal kingdom and the plant kingdom (Burge et al. 1998).

Another distinguishing feature of U12-type introns is that the distance between the branch site adenosine and the acceptor site (DistBA) is unusually short, between 10 bp and 20 bp (Sharp and Burge 1997), while the DistBA of the U2-type introns can be over 100 bp (Smith and Nadal-Ginard 1989). It has also been experimentally confirmed that spacing mutations with the DistBA less than 10 nt or more than 20 nt would strongly activate cryptic 3′ splice sites (Dietrich et al. 2001a). As a result, Dietrich et al. (2001) proposed a local diffusion model to explain the acceptor site selection of the U12-type intron.

The initial recognition of the U12-type introns arose from its non-canonical dinucleotide termini AT-AC (Jackson 1991), distinct from the conventional GT-AG intron borders. Further research indicated that GT-AG introns can be spliced by U12-type spliceosomes, and, conversely, AT-AC introns can be spliced by U2-type spliceosomes (Wu and Krainer 1997; Dietrich et al. 1997). Therefore, intron type cannot be simply determined by the dinucleotide termini. This raises the question of how to distinguish U12-type introns from U2-type introns. Based on conserved motifs of the donor site and the branch site in the U12-type introns, Burge et al. (1998) designed a computer program, named U12Scan, to address the issue of the identification of U12-type introns and conducted a survey in a variety of species based on the GenBank gene structure annotation. Later, Levine and Durbin (2001) adopted a slightly different strategy to recognize human U12-type introns. They predicted U12-type introns in the human genome first, and confirmed the hypothetical introns by expressed sequence data, requiring a 64 bp perfect match between a transcript sequence fragment and the 32 bp flanking sequences of a predicted U12-type intron in both directions. The latter approach did not suffer from the incompleteness or likely errors in the GenBank annotation, but has its own problems. For example, any U12-type intron flanked by exons shorter than 32 bp would not be located. In addition, both analyses restricted their search of DistBA within the region [8, 21], under the presumption that no U12-type introns have DistBAs shorter than 8 nt or longer than 21 nt.

In a recent study, we mapped 176,915 *Arabidopsis* ESTs on the *Arabidopsis* genome, and 45 U12-type introns were identified in the EST-confirmed introns (Zhu et al., 2003). A more sophisticated analysis was undertaken in this study including 26,961 *Arabidopsis* full-length cDNAs in addition to the EST set used in the previous study. A total of 158 distinct U12-type introns were

identified, including 50 AT-AC introns, 1 AT-AA intron and 107 GT-AG introns, comprising many more than the overall number of U12-type introns previously reported in plants. Subsequent analysis indicates that *Arabidopsis* U12-type introns not only share similar features with *Arabidopsis* U2-type introns in intron length distribution and low GC content relative to the flanking exons, but also share almost identical splice signals with U12-type introns from other species. One significant discovery is that 5 U12-type AT-AC introns and 7 U12-type GT-AG introns have DistBAs longer than 21 nt, the longest observed distance being 35 nt. When further extending the BPS search region, another novel U12-type GT-AG intron was identified with a DistBA of 39 bp. The presumed 21 bp maximum limit appears incorrect, even though the distribution of DistBAs of the U12-type introns shows a peak at 12 nt. Several alternative splicing events involving U12-type introns were also found in this study and provide direct evidence that U12-dependent 3'ss could be recognized by the U2-type spliceosome. Analysis of the cases of alternative splicing combined with dinucleotide preference analysis also demonstrates that the sequence context of the potential acceptor site may have an important role in the 3' splice site selection, in addition to the spacing constraint. Furthermore, phylogenetic analysis dates the fusion event of the two spliceosomes as subsequent to the divergence of AtNHX1-4 (*Arabidopsis* Na+/H+ antiporter 1-4) and AtNHX5-6.

MATERIALS AND METHODS

cDNA/EST- confirmed introns in the Arabidopsis thaliana genome

The *A. thaliana* genome sequence (released on Aug. 20, 2002) was retrieved from GenBank (http://www.ncbi.nih.gov/GeneBank/), with accession numbers NC_003070, NC_003071, NC_003074, NC_003075 and NC_003076 for the five chromosomes respectively. *Arabidopsis* full-length cDNA sequences were also downloaded from GenBank (dated 11/04/2002), and *Arabidopsis* ESTs were downloaded from the NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) in December, 2002. All 27,288 putative *Arabidopsis* proteins (data label: ATpep, version: July 25, 2002) were downloaded from The Institute of Genome Research (ftp://ftp.tigr.org/pub/data/a_*thaliana*/ath1/SEQUENCES/ATH1.pep), which represented the annotation of the same *Arabidopsis* genome release as the one used in this study.

A total of 26,961 full-length cDNAs and 176,915 ESTs were aligned with the *Arabidopsis* genome sequence using the GeneSeqer spliced alignment program (Usuka, et al., 2000) at high

stringency in order to generate a reliable dataset of *Arabidopsis* introns. The cDNA/EST-confirmed introns originated from the putative cognate spliced alignments with local similarity scores higher than 0.9 (Zhu, et al. 2003), and qualified introns were merged into a non-redundant intron set for subsequent analysis.

Identification of Arabidopsis U12-type introns

The identification procedure used follows that established by Burge et al. (1998). A simple description is repeated in the following. First, weight matrices for the splice sites of U12- and U2-type introns were derived from subsets of the transcript-confirmed *Arabidopsis* introns. 47 confirmed introns with AT-AC termini and two introns with AT-AA termini were selected as the training set of U12-type introns according to their splice signals. The weight matrices for the recognition of U12-type introns in the subsequent analysis were generated with the MEME program (Bailey and Elkan 1994). In addition, 70,189 cDNA/EST-confirmed GT-AG introns which lack the sequence ATCC in positions +3 to +6 relative to the 5′ splice site (5′ss) were utilized as a training set to construct the corresponding weight matrixes for the U2-type introns. The probabilities of the splice signals were then computed as the products of the corresponding position-specific probabilities, based on the observed residue frequencies derived from the transcript-confirmed introns. The log-odds ratio of the score derived from the U12-type splice signals versus that from the U2-type splice signals was computed for the 5′ss and the BPS of all the transcript-confirmed introns including the training sets. The log-odds ratios were further z-normalized as S_x, where x is d or b, denoting the donor site and the BPS, respectively. Thus, the introns with large values for both S_d and S_b were selected as U12-type introns (see Fig.1; also see Burge et al., 1998). Because one of the U12-type likely AT-AC introns has DistBA as long as 35 nt, we set the search region for the branch-site motif as [-42, -5] relative to the confirmed 3′ splice sites, corresponding to DistBA in the range [6, 35]. The region [-5, +5] relative to the 5′ss was also scanned for possible ambiguities in case the exon-intron junction cannot be determined unambiguously by spliced alignment because of sequence repeats, and the ambiguous cases are thereby corrected after the further confirmation.

Dinucleotide relative abundance in the proximity of the acceptor site of U12-type introns

Let f_x and f_{xy} represent the frequency of the nucleotide x and the frequency of the dinucleotide xy, respectively. The dinucleotide relative abundance is defined as $\rho_{xy}=f_{xy}/f_x f_y$, as a common assessment of dinucleotide bias (Burge, et al., 1992). Dinucleotide relative abundances were calculated for the region between 10 bp downstream of the branchpoint adenosine and one bp upstream of the 3'ss and also in the equal size region immediately downstream of the 3'ss within the exon. As a control, dinucleotide relative abundances were also derived for the U2-type GT-AG intron sequences in the 10 bp regions immediately preceding and succeeding to the 3' terminal dinucleotide. Hence, if the 5'-most AC downstream of the BPS is almost always selected as the 3'ss in U12-type AT-AC introns, the dinucleotide AC should be under-represented between BPS and 3'ss, that is, ρ_{AC} should be sufficiently smaller than 1.

Gene Duplications

A recent study suggested that the *Arabidopsis* genome has undergone at least two large-scale duplications and identified 3,044 gene pairs divided into 91 chromosomal blocks (Blanc et al. 2003). The authors concluded that one event was a recent polyploidy which occurred 24-40 mya (millions years ago) and that the other event was an older one which happened after the monocot/dicot divergence. The 3,044 gene pairs and the related information were downloaded from http://wolfe.gen.tcd.ie/athal/dup and used to study the fate of the U12-type introns after gene duplication.

RESULTS

Identification and Characteristics of U12-dependent introns

There are 53 introns with AT-AC terminal dinucleotides and six introns with AT-AA termini in the non-redundant transcript-confirmed *Arabidopsis* intron set that were identified as candidate U12-

type introns. Because of the absence of the typical motifs for both the donor site and the branch site, six AT-AC introns and four AT-AA introns were removed. The remaining 49 introns were utilized to build weight matrices for 75,717 transcript-confirmed introns were computed (see Methods), projecting these introns into points in the two-dimensional plane (Fig. 1). As expected, the 49 U12-type like AT-AM introns from the training set map in the upper-right corner in the plot, accompanied by one GT-AT introns and hundreds of GT-AG introns. Consistent with the manual inspection mentioned above, six AT-AC introns and four AT-AA introns map close to the origin in the plane and thus predicted to be spliced by the U2-type spliceosome. One AT-AC intron and one AT-AA intron included in the training data set have relatively low values in either S_d or S_b when compared to the other 47 U12-type AT-AM introns (enclosed in the yellow rectangle in Fig. 1). We conservatively excluded these two introns in the further analysis. The remaining 47 AT-AM introns were selected as authentic U12-type introns for reference. Because there is not an obvious cluster to separate the putative U12-type introns from U2-type introns, the determinant of U12-type introns versus U2-type introns was empirically defined with respect to the standardized scores of the 47 introns in the reference set, such that the U12-type intron should satisfy the following conditions: S_d and S_b are no less than the minimum value of S_d and S_b from the 47 U12-type AT-AM introns, respectively. The qualified introns roughly enclosed by the yellow rectangle in Fig. 1 include 110 GT-AG introns, 46 AT-AC introns, one AT-AA intron and one GT-AT intron.

All 158 predicted U12-type introns and related information are listed in Table 1, together with another four U12-type AT-AC introns identified by a BLASTP search (Altschul et al. 1997) described in the next section. As shown in Table 1, there are four U12-type GT-AG introns that are alternatively spliced with cryptic acceptor sites in the proximity of the normal 3'ss, which leads to ambiguity in the DistBA. Thus, the analysis of the DistBA distribution was based on the 154 distinct introns after excluding the four pairs of introns involved in alternative splicing. As shown in Fig. 2, the DistBA distribution of the U12-type AT-AM introns seems similar to that of the U12-type GT-AG introns. In particular, both distributions have the mode at 12 nt, and in both sets the shortest DistBA is 11 nt. Interestingly, 12 U12-type introns (five AT-AC introns and seven GT-AG introns) have distances longer than 21 nt, the maximum distance previously reported (Levine and Durbin 2001; Dietrich et al. 2001a). To be conservative in assessing the authenticity of the U12-type introns in this study, the introns with long DistBA (>21 bp) were left out in the subsequent analysis of the sequence characteristics of U12-type introns.

Thus, 51 U12-type AT-AM (including the four novel U12-type AT-AC introns identified in the next section) and 103 U12-type GT-AG introns are under further analysis. The *Arabidopsis* U12-dependent splice signals display similar base composition to that of previously identified U12-type introns from various species (Burge et al. 1998). There is no significant difference in the length distribution between the U12-type introns and U2-type introns in *Arabidopsis* (Fig. 3), in contrast to reported lack of short U12-type introns relative to U2-dependent introns in human (Levine and Durbin 2001). Furthermore, plant introns are characterized by low GC content when compared to the flanking exons (Goodall and Filipowicz, 1989), and our analysis shows that U12-type introns have this trait in common with the U2-type introns in *Arabidopsis* (data not shown).

Gene Duplications and Molecular Phylogeny Analysis

Of 3,044 pairs of duplicated *Arabidopsis* genes (Blanc et al. 2003), 24 pairs of genes have at least one U12-type intron in one or the other gene (Table 2). Based on our stringent criteria, the candidate U12-type introns are highly likely to be authentic U12-type introns, but the remaining transcript-confirmed introns may not necessarily be spliced by the major spliceosome. On the other hand, lack of transcript evidence may also cause some U12-type introns not to be identified either. Thus, introns paired with U12-type introns were examined and thereby three U12-type AT-AC introns and three U12-type GT-AG introns. Of the remaining unidentified introns, two U12-type introns have no paralogous introns in the paired genes, presumably due to intron loss/gain. In addition, two GT-AG introns lost U12-specific splice signals and are recognized as U2-type introns, whereas the classification of five GT-AG introns is not certain because their S_d and S_b scores are big but not large enough to satisfy the criteria. There are only two gene pairs derived from ancient large-scale gene duplications as listed in Table 2, and the two cases of U12-type GT-AG introns converted into U2-type GT-AG exactly come from the two pairs. Excluding the five ambiguous cases, 15 of the remaining 19 pairs U12-type introns were stably conserved since the divergence about 24-40 mya (Blanc et al. 2003). Thus, U12-type introns seem to be very stable in recent gene duplication, but are likely to be converted into U2-type introns in the long run.

Because occasional (random) gene duplications occur in addition to the large-scale segmental genome duplications, we searched all the genes containing U12-type AT-AM introns against ATpep using BLASTP, and thus identified another novel U12-type AT-AC intron in the gene At1g76170 based on the non-cognate EST gi:23303104 from its paralogous gene At2g44270. The BLAST

search also revealed more cases of the conservation of U12-type introns. First, the gene At1g79610, which encodes a low abundance Na+/H+ antiporter (AtNHX5) in shoots and roots in *Arabidopsis* (Kmieciak et al. 2002), contains a total of two U12-type AT-AC introns, one U12-type GT-AG intron and 17 U2-type introns (Table 1). AtNHX5 shares it's a highly conserved sequence and gene structure with another family member, AtNHX6 (from gene At1g54370), which has two corresponding U12-type AT-AC introns (listed in Table 1). The intron type of the counterpart in AtNHX6 of the U12 GT-AG intron of AtNHX5 is uncertain, because the intron has a strong U12-dependent donor site but a relatively weak U12-dependent branch signal (TTCATGAC) with an 11 bp DistBA (indicated by the yellow hollow arrow in Fig. 1). However, the intron has a strong U12-dependent branchpoint signal (TCCTTGAC) with a 39 bp DistBA, whereas the branch site of the corresponding U12-type intron in AtNHX5 has a DistBA of 32 bp. Whether the intron is an authentic U12-type intron and whether the high score branch site is functional in *in vivo* may have to be determined by experimental methods. It would be the longest DistBA for the U12-type introns identified to date, if confirmed. There are four other members of Na+/H+ antiporter in *Arabidopsis* (AtNHX1-4), and AtNHX5 and AtNHX6 have more sequence similarity with the human Na+/H+ exchangers HsNHE6 and HsNHE7 (Kmieciak et al. 2002). Interestingly, there are two U12-type GT-AG introns and one U12-type AT-AC intron in HsNHE6 (Levine and Durbin 2001). Further analysis indicates that there are no U12-type introns in AtNHX1-4 and HsNHE7, and the latter have completely different gene structures when compared to AtNHX5-6 and HsNHE6, respectively. With respect to the cladogram of Na+/H+ antiporters from *Arabidopsis*, human, rice, *E. coli*, yeast, and other species (see Fig. 2 in Kmieciak et al. 2002), we can infer that the appearance of the U12-type introns is dated prior to the divergence of AtNHX5-6 from HsNHE6-7, but after the divergence of AtNHX5-6 and AtNHX1-4.

As a second example, the genes At1g02750, At1g56280, At3g05700, At3g06760, At4g02200, At5g26990 and At5g49230 that encode drought-induced like proteins, all have one U12-type AT-AC intron in the same location. The only exception is the gene At4g02200, which has a U12-type GT-AG intron instead. After correcting annotation errors based on the transcript sequence data, the protein sequences of the seven genes and a homologous gene (accession number AA033770) from rice were aligned by ClustalX (Fig. 4A; for ClustalX see Shukla et al. 2002) and a neighbor-joining tree was constructed based on the multiple alignment using MEGA2.1 (Fig. 4B; for MEGA2.1 see Kumar et al. 2001). Detailed analysis indicates that the gene structures are highly conserved among the seven *Arabidopsis* genes and the rice gene. The identified U12-type introns are all in coding

phase 0 and the same position starting after the conserved lysine (K103, highlighted in green in Fig. 4A), where a U2-type GT-AG intron is located in the rice homolog. We were particularly interested in this example to find out how the U12-type AT-AC intron switches to the U12-type GT-AG intron in the gene At4g02200 since the divergence from the gene At1g02750. The multiple alignment of At4g02200, At3g05700 and At1g02750 suggests that the conversion was probably initiated by the mutation of the 5' terminal AT to GT, with subsequent activation of an AG downstream of the original acceptor site as the canonical 3'ss (Fig. 4C).

The BLASTP similarity search increased the number of identified U12-type AT-AC introns by four, suggesting that paralogous transcripts could also be very helpful in the identification of U12-type introns. And, the high conservation of the U12-type introns among paralogous genes were also observed in other studies (Burge, et al., 1998; Levine and Durbin, 2001).

Alternative splicing

Seven cases of alternative splicing events related with U12-type introns were captured in this analysis. Four of them, as highlighted in Table 1, involve alternatively activated cryptic acceptor sites in the proximity of the normal splice sites. In detail, the U12-type introns in the genes At2g26430 and At3g13460 both have minor isoforms utilizing CAG/, seven and nine nt downstream of the cognate acceptor site TAG/, respectively (where / denotes the exon-intron junction). It suggests that the distal AG with the favored sequence motif may compete with the first AG downstream of the BPS in the selection of the acceptor site of U12-type GT-AG introns. The "leaky" scan revealed by the two alternative splicing events also implies that the spacing constraint might not be as strong as presumed in the 3'ss selection in the U12-type intron splicing. Differing from the previous two examples, the U12-type GT-AG intron in At3g52180 has a cryptic acceptor site 28 nt upstream of the wild type acceptor and 14 bp preceding the normal U12-dependent branch site. The motif search of U12-type branchpoint signals indicates that the most "likely" branch site signal is TCCTTCGC with the DistBA 28 bp. However, it is uncommon for a guanosine to be bulged at the branch site. It is likely that the upstream GTTTTCAC is employed in splicing with the DistBA of 38 bp. The alternative explanation is that the isoform might be spliced by the U2-type spliceosome rather than the minor spliceosome. The U12-type intron in the gene At4g09720 has a cryptic and unusual acceptor site AT/ (five nt ahead of the cognate 3' splice site) with the DistBA of 10 nt which is the shortest DistBA found in this study.

Of the remaining three instances of alternative splicing not listed in Table 1, one alternatively utilized the wild-type donor site /GC (with six transcript evidences; see supplemental material) three nt preceding to the cryptic U12-dependent donor site /GT (with two transcript evidences) in the gene At2g44680. Further analysis revealed that the U12-type GT-AG intron in the paralogous gene At3g60250 is also alternatively spliced with the cryptic 5'ss /GC (with one transcript evidence) three nt prior to the cognate U12-type donor site /GT (with three transcript evidences). It is of particular note that the major isoform is the U12-type 5'ss /GT in the gene At3g60250 and becomes the U2-type 5'ss /GC in the gene At2g44680. The last example is an exon skipping event in the gene At1g49160, a putative serine/threonine protein kinase. The U12-type GT-AG intron in At4g160 is alternatively spliced using the donor site of the upstream U2-type GT-AG intron, whereas there is no evidence to confirm whether the exon skipping also occurs in the U12-type GT-AG intron in its paralog (the gene At3g18750, listed in Table 2). The above three intron isoforms all have U2-type donor site and are likely to be spliced by the U2-type spliceosome, suggesting that U12-type 3'ss can also be recognized by the major spliceosome. These examples also reveal a potential pathway for the conversion from U12-type GT-AG introns to U2-type introns.

Additionally, retention of the U12-type AT_AC intron was recorded in the genes At1g73350 and At5g63700. The alternative intron retention might be a step preceding the loss of the U12-type AT-AC introns.

Selection of the acceptor site of U12-type introns

It was noted that only four combinations of the terminal dinucleotides were observed in this large scale analysis on the *Arabidopsis* genome, and GT-AG and AT-AC take the majority while only one U12-type AT-AA intron and one U12-type GT-AT intron were identified (the latter belongs to a splicing isoform as mentioned above). It seems that the selection of the 3'-terminal dinucleotides of U12-type introns is highly correlated with the selection of the 5'-terminal dinucleotides of U12-type introns, that is /GT is typically matched with AG/, and /AT paired with AC/ or AA/ occasionally. In order to find out the whether the scanning model is applicable in the selection of the acceptor site of U12-type introns, the dinucleotide preference was computed for the regions in the proximity of the U12-type acceptor site (see Methods; also see Fig. 5). The results indicate that all dinucleotides

starting with adenosine are underrepresented prior to the U12-dependent acceptor site, and it is interesting that AC is not the least preferred among U12-type AT-AC. Further analysis revealed that there is a total of six AC occurrences between the BPS and 3'ss in the total 51 AT-AM introns. All six AC are located in introns with DistBA less than 18 nt, each of them is immediately prior to the 3'ss AC/, and not one contains C immediately upstream. Thus, CAC/ is strongly preferred as the acceptor site of U12-type AT-AC introns in addition to the DistBA constraint. It seems that the selection of the U12-dependent acceptor site neither follows a simple scanning mechanism (that is, the first AC following the branch point is selected as the acceptor site). It is likely that the sequence surrounding the 3' splice site also plays an important role in the selection of U12-dependent acceptor sites, and the selection might be mediated via exon definition. In addition, the five U12-type AT-AC introns with DistBA larger than 21 nt all have CAC/ as the acceptor site, further indicating the importance of the surrounding sequence in the selection of U12-type 3' splice site.

AG is strongly avoided in the proximal region prior to the donor site in either of the two classes of introns, but is a little more preferred in the exon regions immediately succeeding to the U12-type introns versus the U2-type introns. Excluding the U12-type introns involved with alternative splicing, only one dinucleotide AG immediately prior to the cognate 3'ss CAG/, in the gene At4g02200, is found in the similarly defined search region for the U12-type GT-AG introns. It seems that the scan model is more applicable to the U12-type GT-AG introns than to the U12-type AT-AC introns.

DISCUSSION

Identification of U12-type introns

In this study, we identified a total of 162 U12-type introns, including one AT-AA, 50 AT-AC, 110 GT-AG and one GT-AT introns by following the procedure proposed by Burge et al. (1998). Burge et al. also provided the test statistics $t = S_b^2 + S_d^2$ in their study, to discriminate U12-type introns against U2-type introns, such that introns with t-scores higher than 20 are likely to be U12-type introns. This criterion implies that an intron with a strong U12-dependent donor site signal (i.e., $S_d > \sqrt{20} = 4.47$) will still be spliced by the U12-type spliceosome, in spite of the weak branch site signal. This may be reasonable on the basis of experimental evidence which revealed that the donor site and the branch site are simultaneously interacting in a U11/U12 di-snRNA complex in the initial

recognition of the U12-type intron by the minor spliceosome (Frilander and Steitz 1999). However, this scheme is not consistent with the observation that the normal U12-dependent intron splicing was abolished by the mutations of the BPS (Dietrich et al. 2001b). Therefore, a more conservative criterion was set for the minimum cut-off values of S_d and S_b in this study, with respect to the U12-type AT-AM introns. Compared to *Arabidopsis* U12-type introns reported in the earlier study (Burge et al., 1998), 10 out of 11 were also recovered in this study with the exception of one AT-AA intron (represented by the green upside-down triangle below the yellow rectangle in Fig. 1) in the gene G5p. On the other hand, introns with high prediction scores might be actually excised by the major spliceosome. For example, the U12-type intron in the gene At3g52180 alternatively activates the cryptic acceptor site 28 bp prior to the wild-type acceptor site, and the predicted BPS is TCCTTCGC with the normalized score 1.66, slightly higher than the minimum requirement. However, there is no adenosine in the sequence TCCTTCGC, therefore it is more likely that the "weaker" BPS GTTTTCAC in the upstream is utilized or the intron isoform is alternatively spliced by the U2-type spliceosome. On the whole, the statistical significance may not be necessarily equivalent to the biological significance. Our predictions, even under stringent criteria, may still need additional biological experiments to confirm, and a more appropriate method will come out for the U12-type intron recognition with the accumulating experimental data in the future. Nonetheless, this research will enrich our understanding of the U12-type introns in *Arabidopsis*, and may also shed light upon studies on the U12-type introns in other species.

Characteristics of Arabidopsis U12-dependent introns

The identified *Arabidopsis* U12-dependent introns display patterns almost identical to the motifs of the U12-type introns from various other species (data not shown), which is in accordance with the postulated early common origin of U12-type introns predating the divergence of animals and plants (Wu et al. 1996; Burge et al. 1998). Different from the characteristics of the U12-type introns recently identified from the human genome (Levine and Durbin 2001), however, our results illustrate that there is neither an appreciable difference in the distribution of the intron length between the U2-type intron and the U12-type intron in *Arabidopsis*, nor is there a significant difference in the distribution of the DistBAs between the U12-type AT-AM introns and U12-type GT-AG introns (see Fig. 2 and 3). Such contrast may arise from the organism difference between humans and *Arabidopsis* or be caused by differences in the analysis method and criteria as discussed above.

The distance between the branch site and the 3′ splice site and the selection of the acceptor site of U12-type introns

Both of the two previous large scale computational scans for U12-type introns were restricted to DistBA between 8 bp and 21 bp (Burge et al. 1998; Levine and Durbin 2001), based on the distribution of the DistBAs of naturally occurring U12-type introns (Sharp and Burge 1997). Experimental evidence mainly came from the recent study on the spacing mutants of the human P120 gene, in which the unfavorable dinucleotide UU with the DistBA of 12 nt was selected as the 3′ splice site rather than the downstream AC with the DistBA of 27 nt as demonstrated by the construct +27 AC (Dietrich et al. 2001a). However, the conclusion that the DistBA constraint is extremely strong in U12-type intron splicing seems less convincing, because the uncommon guanosine immediately prior to the +27 AC might disable the dinucleotide as a functional acceptor site in that construct. Also, our results indicate that this model is not generally valid, or is at least not valid in *Arabidopsis*, based upon three observations. First, at least 12 U12-type introns have DistBAs larger than 21 bp, even though the mode of the DistBA distribution is 12 nt (see Table 1 and Fig. 2). Second, only one AT-AA intron and one GT-AT intron were found in addition to the U12-dependent GT-AG and AT-AC introns, suggesting that the combination of GT-AG, or AT-AC is strongly preferred in nature among U12-type introns. Each of the two observations indicates that the DistBA constraint is not as strong as the preference for the terminal dinucleotide combination. Finally, as mentioned above, there are six cases in which the dinucleotide AC is located immediately prior to the confirmed 3′ splice sites in the U12-type AT-AC introns, indicating that the 3′ss surrounding sequence also plays an important role in the selection of the acceptor site of U12-denpendent introns. Similar results were also observed in the U12-type GT-AG introns. Furthermore, the alternative 3′ss events in the U12-type GT-AG introns also confirm that distal acceptor sites with the favored sequence can compete with the proximal wild type 3′ss in U12-type intron splicing.

A caveat concerning the existence of long DistBAs is that there might be a weak BPS actually functioning in the downstream of the most likely predicted BPS. Another possible argument is that the introns with long DistBA (>21 bp) might actually be spliced by the U2-type spliceosome. Neither of the two arguments has any experimental support thus far, therefore it is more reasonable to assume that a small number of U12-type introns have longer DistBAs than previously thought.

This conclusion results in a new question: does any U12-type intron have a DistBA longer than 35 nt, out of the searching region [6, 35] in this study? To address this issue, we extended the searching region from 35 bp to 45 bp DistBA, to see whether new U12-type introns were discovered. Only one intron in AtNHX6 (the gene At1g54370) has a stronger BPS with the presumptive DistBA as 39 nt, as discussed above. Nevertheless, there are few introns with a DistBA longer than 35 nt in *Arabidopsis*, and the upper limit may be easily tested experimentally.

High conservation was observed in the functional regions of U6atac and U12 snRNAs between human and *Arabidopsis* (Shukla and Padgett 1999), so we believe that the long DistBA observed in the *Arabidopsis* U12-type introns in this study is likely to also apply to humans or to other species. Hence, rescanning the genomes of humans and other organisms with an extended searching region may reveal more novel U12-type introns.

Evolutionary origins and fates of U12-type introns

To date, the fission/fusion model proposed by Burge et al. (1998) is well accepted for the origin of the U12-dependent spliceosome and the excised introns. For the first time, the phylogenetic analysis of the Na+/H+ antiporter family dated the likely fusion event as being prior to the divergence of the plant kingdom and the animal kingdom and subsequent to the divergence of AtNHX1-4 and AtNHX5-6. It also gives the first evidence that the U12-type introns are evolutionarily stable over one billion years. This amazing stability likely results from the unusual conserved U12-dependent 5'ss and BPS, so that any mutation in the splice signal sequences may easily disrupt the normal splicing of the U12-type AT-AC introns and thereby is strongly selected against. The alternative splicing examples, however, also demonstrate that U12-type AT-AC introns can be lost by intron retention, which is probably caused by mutations in the U12-dependent splice signals. In addition to the experimental evidence indicating that the U12-dependent 5'ss can be exploited by the major spliceosome (Dietrich et al., 1997), the alternative splicing events captured in this study also indicate that the 3'ss (probably as well as the BPS) of the U12-type GT-AG introns can also be utilized the major spliceosome. Therefore, mutations that corrupt the conserved U12-dependent splice signals may easily trigger the conversion from the U12-type GT-AG introns to the U2-type GT-AG intron, while the reverse process is highly improbable.

Besides stability or loss, the fate of U12-typte AT-AC may also involve switch to U12-type GT-AG introns. A plausible mechanism of the switch is as follows. The 5′ terminal dinucleotide /AT mutates to /GT and then the first AG or the distal AG with the favored surrounding sequence in the downstream of the BPS is selected in the U12-type intron splicing. Under this model, the mutation is likely to cause the downstream exon to be truncated or extended with the broken reading frame. Another mechanism of the switch proposed by Burge et al. (1998) is from AT-AC to AT-AG and then to GT-AG. However, because it requires two mutation events and the occurrence of natural AT-AG introns is extremely low, the pathway seems not plausible either. At any rate, the switch between U12-type AT-AC intron and U12-type GT-AG intron should be much rarer than the conversion from the U12-type GT-AG intron to the U2-type GT-AG intron. However, it is difficult to explain why the U12-type GT-AG introns outnumber the U12-type AT-AC introns. We are forced to infer that U12-type GT-AG introns did not originate from U12-type AT-AC introns but appeared together with the latter 1 billion years ago. Another more plausible explanation is that there is some kind of selection against the conversion from U12-type GT-AG to U2-type GT-AG intron. It was noted that most of the genes containing U12-type introns function in information processing (Burge et al. 1998). In this study, some of the genes containing U12-type introns were found to be stress reaction related, such as the drought-induced like proteins (Fig. 4) and AtNHX5, whose expression level increases in response to salt treatment (Yokoi et al. 2002). It is possible that the U12-dependent spliceosome system might be activated and therefore regulates the expression level of target genes which contain U12-type introns via changing the speed of U12-type intron splicing (Patel et al. 2002) in response to stresses in plants or the analogous situations in vertebrates or insects. In this scheme, the potential role of the U12-dependent spliceosome system may result in selective pressure against the conversion from U12-type GT-AG introns to U2-type GT-AG introns.

In spite of the high stability and possible selective advantage, it is likely that the number of U12-type introns has been slowly but continuously reduced by accumulating mutations. Gene duplication, however, may help the U12-type introns propagate within the gene families, for example, the drought-induced like protein family and dTDP-glucose 4-6-dehydratease like proteins (Zhu et al. 2003).

SUPPLEMENTARY MATERIAL

All spliced alignments are interactively accessible at AtGDB (Arabidopsis thaliana genome database, http://www.plantgdb.org/AtGDB/).

ACKNOWLEDGEMENTS

The authors would like to thank Jacqueline E. Townsend for the preparation of the manuscript and Richard A. Padgett for kindly providing reprints. This work was supported in part by NSF grant DBI-0110254 to V.B.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-402.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28-36.

Blanc, G., Hokamp, K., and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* **13:** 137-44.

Burge, C., Campbell, A.M., and Karlin, S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci U S A* **89:** 1358-62.

Burge, C.B., Padgett, R.A., and Sharp, P.A. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell* **2:** 773-85.

Burge, C.B., T. Tuschl, and P.A. Sharp. 1999. "Splicing of precursors to mRNAs by the spliceosome." In R.F. Gestland, T. Cech, and J.F. Atkins, editors, *The RNA world II*. Cold Spring Harbor Laboratory Press . Cold Spring Harbor, N.Y. p.525-560.

Dietrich, R.C., Incorvaia, R., and Padgett, R.A. 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell* **1:** 151-60.

Dietrich, R.C., Peris, M.J., Seyboldt, A.S., and Padgett, R.A. 2001a. Role of the 3' splice site in U12-dependent intron splicing. *Mol Cell Biol* **21:** 1942-52.

Dietrich, R.C., Shukla, G.C., Fuller, J.D., and Padgett, R.A. 2001b. Alternative splicing of U12-dependent introns in vivo responds to purine-rich enhancers. *RNA* **7:** 1378-88.

Frilander, M.J. and Steitz, J.A. 1999. Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes Dev* **13:** 851-63.

Goodall, G.J. and Filipowicz, W. 1989. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* **58:** 473-83.

Hall, S.L. and Padgett, R.A. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol* **239:** 357-65.

Hall, S.L. and Padgett, R.A. 1996. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science* **271:** 1716-8.

Hastings, M.L. and Krainer, A.R. 2001. Functions of SR proteins in the U12-dependent AT-AC pre-mRNA splicing pathway. *RNA* **7:** 471-82.

Jackson, I.J. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res* **19:** 3795-8.

Kmieciak, M., Simpson, C.G., Lewandowska, D., Brown, J.W., and Jarmolowski, A. 2002. Cloning and characterization of two subunits of Arabidopsis thaliana nuclear cap-binding complex. *Gene* **283:** 171-83.

Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244-5.

Levine, A. and Durbin, R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res* **29:** 4006-13.

Luo, H.R., Moreau, G.A., Levin, N., and Moore, M.J. 1999. The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes. *RNA* **5:** 893-908.

Patel, A.A., McCarthy, M., and Steitz, J.A. 2002. The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J* **21:** 3804-15.

Schneider, C., Will, C.L., Makarova, O.V., Makarov, E.M., and Luhrmann, R. 2002. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol* **22:** 3219-29.

Sharp, P.A. and Burge, C.B. 1997. Classification of introns: U2-type or U12-type. *Cell* **91:** 875-9.

Shukla, G.C., Cole, A.J., Dietrich, R.C., and Padgett, R.A. 2002. Domains of human U4atac snRNA required for U12-dependent splicing in vivo. *Nucleic Acids Res* **30:** 4650-7.

Shukla, G.C. and Padgett, R.A. 1999. Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. *RNA* **5:** 525-38.

Shukla, G.C. and Padgett, R.A. 2001. The intramolecular stem-loop structure of U6 snRNA can functionally replace the U6atac snRNA stem-loop. *RNA* **7:** 94-105.

Smith, C.W. and Nadal-Ginard, B. 1989. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell* **56:** 749-58.

Tarn, W.Y. and Steitz, J.A. 1996a. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84:** 801-11.

Tarn, W.Y. and Steitz, J.A. 1996b. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273:** 1824-32.

Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16:** 203-11

Will, C.L., Schneider, C., MacMillan, A.M., Katopodis, N.F., Neubauer, G., Wilm, M., Luhrmann, R., and Query, C.C. 2001. A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J* **20:** 4536-46.

Will, C.L., Schneider, C., Reed, R., and Luhrmann, R. 1999. Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science* **284:** 2003-5.

Wu, H.J., Gaubier-Comella, P., Delseny, M., Grellet, F., Van Montagu, M., and Rouze, R. 1996. Non-canonical introns are at least 10(9) years old. *Nat Genet* **14:** 383-4.

Wu, Q. and Krainer, A.R. 1996. U1-mediated exon definition interactions between AT-AC and GT-AG introns. *Science* **274:** 1005-8.

Wu, Q. and Krainer, A.R. 1997. Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA* **3:** 586-601.

Wu, Q. and Krainer, A.R. 1999. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol Cell Biol* **19**: 3225-36.

Yokoi, S., Quintero, F.J., Cubero, B., Ruiz, M.T., Bressan, R.A., Hasegawa, P.M., and Pardo, J.M. 2002. Differential expression and function of Arabidopsis thaliana NHX Na+/H+ antiporters in the salt stress response. *Plant J* **30:** 529-539.

Zhu, W., Schlueter, S.H., and Brendel, V. 2003. Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiology* **in press.**

Figure Legends

Figure 1. *Identification of U12-type introns.* Each transcript-confirmed intron is projected into the two-dimensional plane with the coordinate (S_d, S_b) in this figure (see Methods), that is approximately distributed as a standard bivariate normal distribution (Burge et al., 1998). The determination of U12-type introns versus U2-type introns is empirically defined with respect to the standardized scores

of the U12-type AT-AC introns, which are clustered in the upper-right corner. The qualified introns are enclosed by the yellow rectangle in the figure. In addition, a yellow arrow indicates a U12-type likely GT-AG intron not included in the selection (see text for the details).

Figure 2. Branch site to acceptor site distance of U12-type introns. The U12-type introns contained in the genes At2g26430, At3g13460, At3g52180 and At4g09720 are not included in this analysis to avoid the uncertainty caused by alternative splicing (Table 1). The remaining 51 U12-type AT-AC or AT-AA introns (black bars) and 103 U12-type GT-AG introns (gray bars) listed in the Table 1 are contributed to this analysis. It is of particular note that 12 U12-type introns (5 AT-AC introns and 7 GT-AG introns) have branch site to 3' splice site distances that are longer than 21 bp.

Figure 3. Length distribution of the U12- and U2-type introns. The density of the length of U2-type introns was derived from 70,189 transcript-confirmed *Arabidopsis* introns (plotted in green line). The histogram of the 154 U12-type introns is represented by the filled columns (see Fig. 2 legend for the details of U12-type intron set).

Figure 4. Drought-induced like proteins. A) Alignment of the protein sequences from *Arabidopsis* and rice. There is a phase 0 intron conservatively located immediately after the green colored column K103 in all of the genes. At that location, the rice gene AA033770 has a U2-type GT-AG intron and the *Arabidopsis* gene At4g02200 has a U12-type GT-AG intron, whereas the remaining 6 genes each have a U12-type AT-AC intron. B) Neighbor-joining tree derived from the alignment in the part A. The branches are colored in green, yellow, and red, for the U12-type AT-AC intron, U12-type GT-AG intron and U2-type GT-AG intron, respectively. C) Alignment of the U12-type intron sequences in the genes At4g02200, At1g02750 and At3g05700. Only terminal alignments are displayed, and the splicing signals of those introns are highlighted in shadow.

Figure 5. Dinucleotide relative abundances in the proximity of the 3' splice site of U12- and U2-type introns. The dinucleotide relative abundance between the branchpoint sequence and the acceptor site versus the equivalent size region immediately succeeding to the acceptor site were plotted for U12-type AT-AM introns (red fonts with underline), U12-type GT-AG introns (green fonts with underline) and U2-type GT-AG introns (blue fonts). The details were described in Methods.

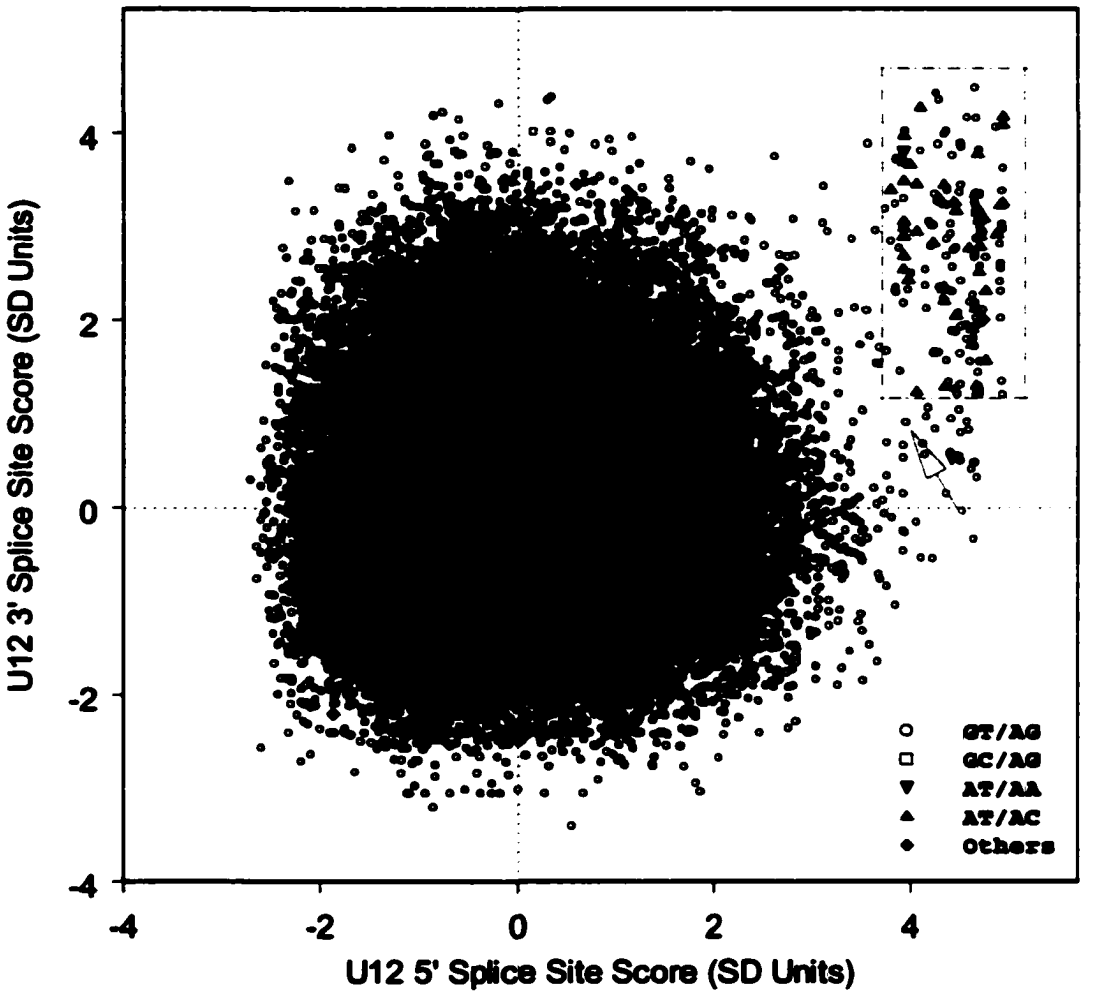

U12 3' Splice Site Score (SD Units)
U12 5' Splice Site Score (SD Units)
4
2
0
-2
-4
-4
-2
0
2
4
GT/AG
GC/AG
AT/AA
AT/AC
Others

Figure 1

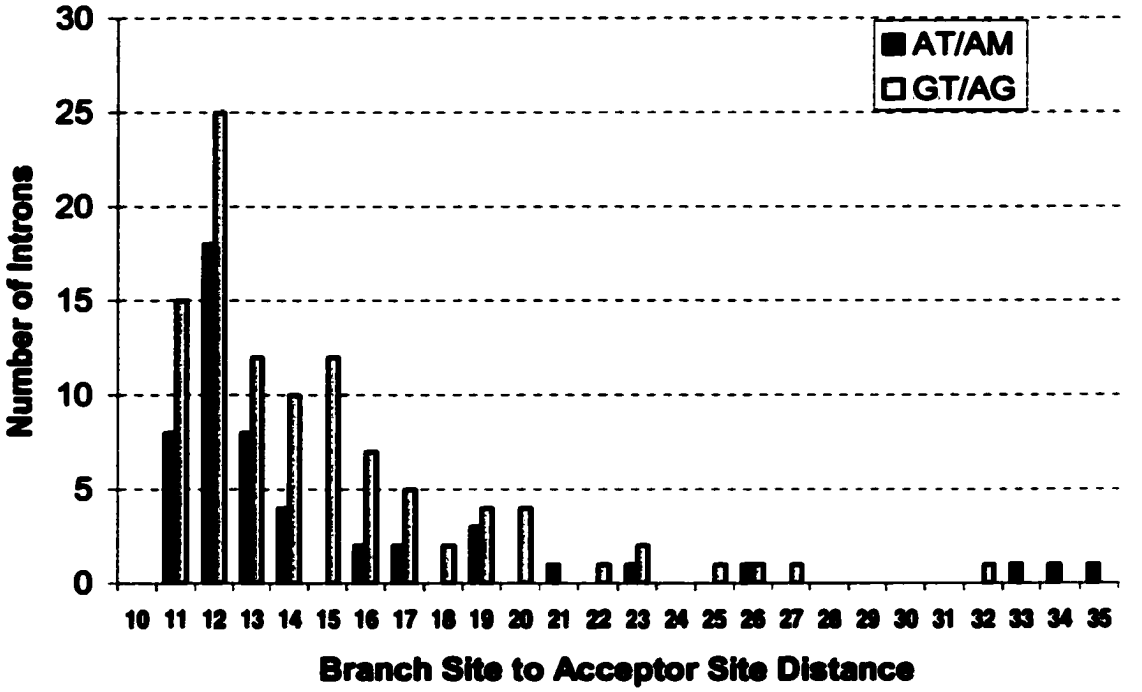

30
25
20
15
10
5
0
Number of Introns
AT/AM
GT/AG
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Branch Site to Acceptor Site Distance

Figure 2

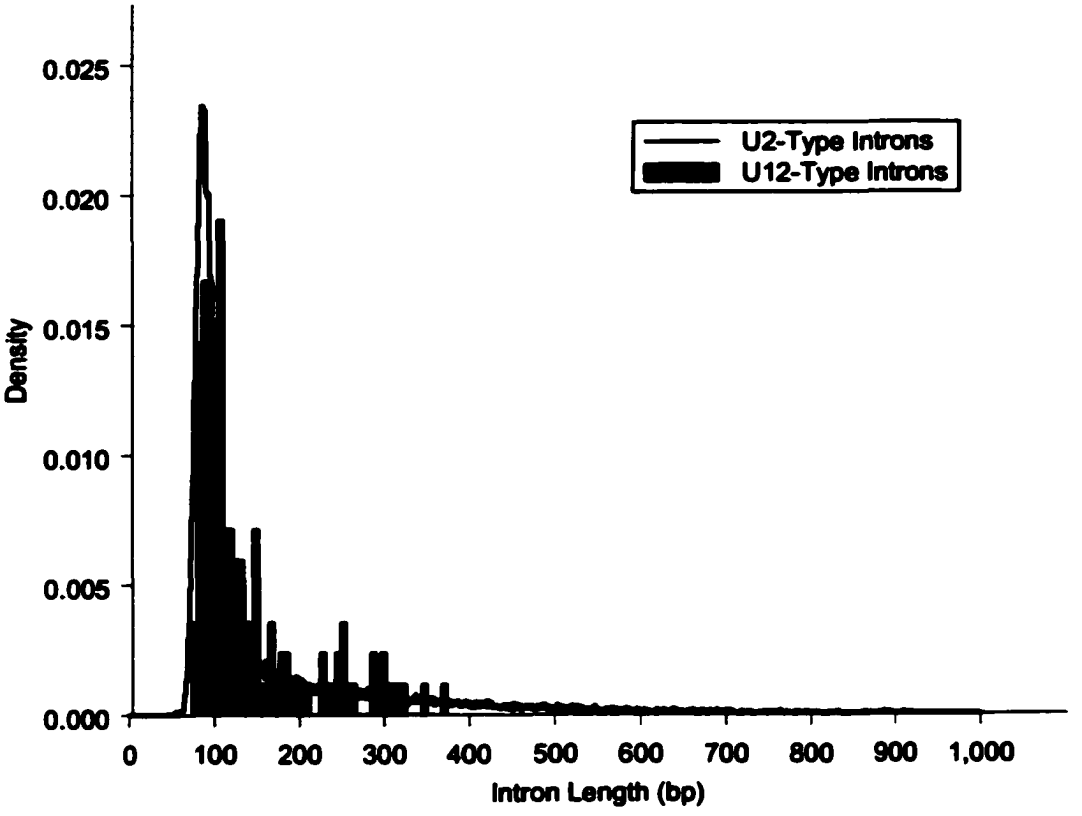
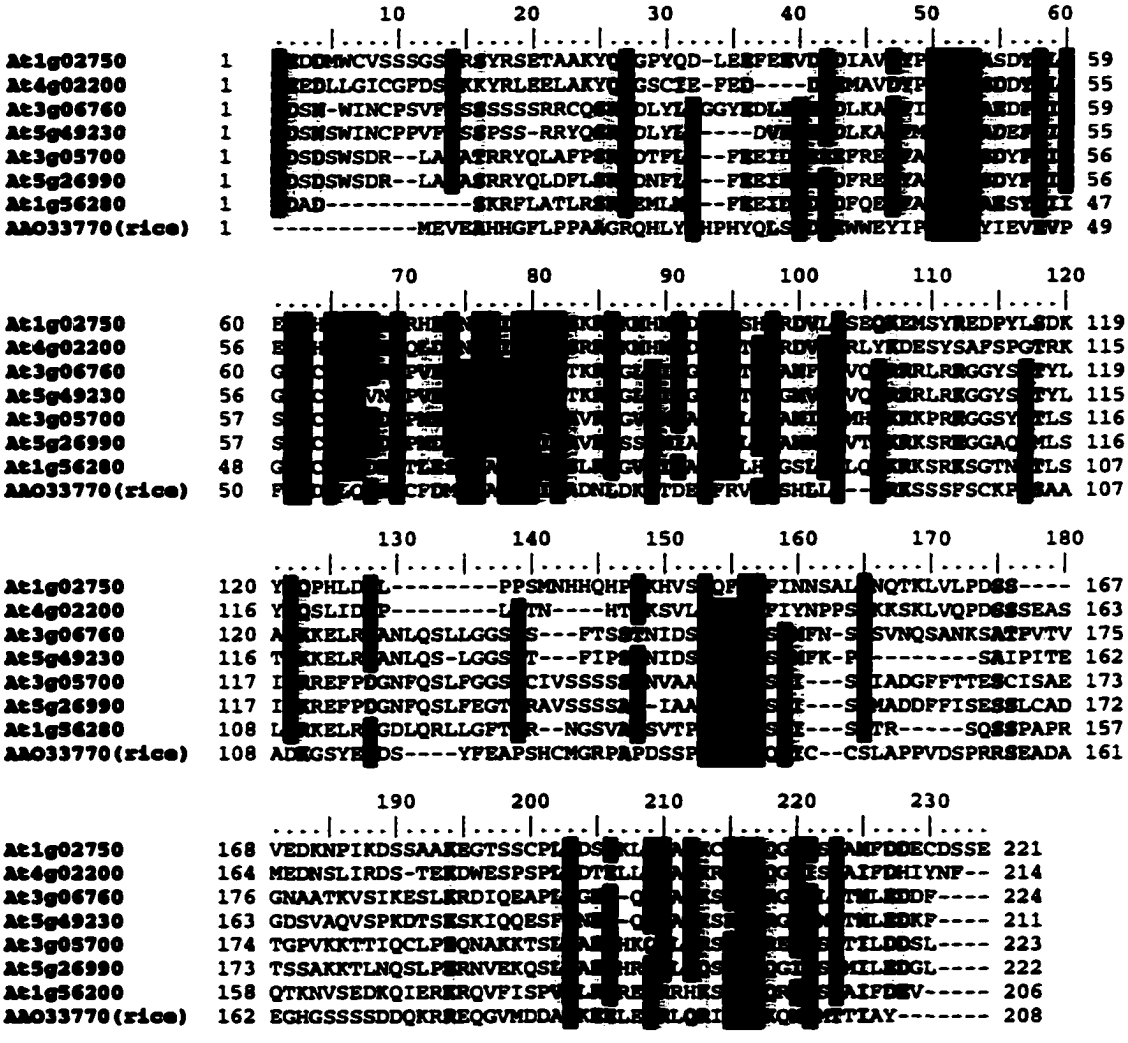

0.025
0.020
0.015
0.010
0.005
0.000
Density
U2-Type Introns
U12-Type Introns
0
100
200
300
400
500
600
700
800
900
1,000
Intron Length (bp)

Figure 3

A.


At1g02750
At4g02200
At3g06760
At5g49230
At3g05700
At5g26990
At1g56280
AAO33770(rice)

B.

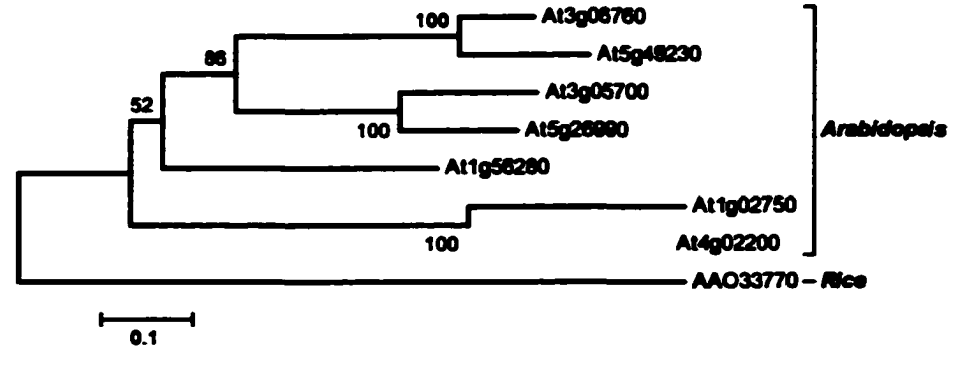

100
At3g06760
At5g49230
86
At3g05700
52
100
At5g26990
Arabidopsis
At1g56280
At1g02750
100
At4g02200
AAO33770 – Rice
0.1

Figure 4

C.

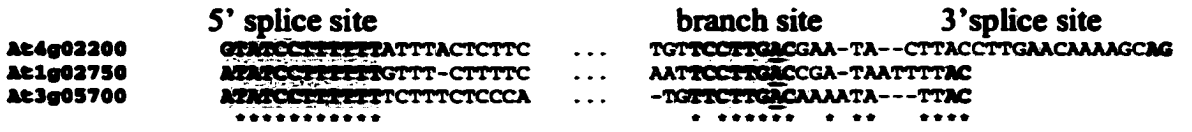


Figure 4 (Cont.)

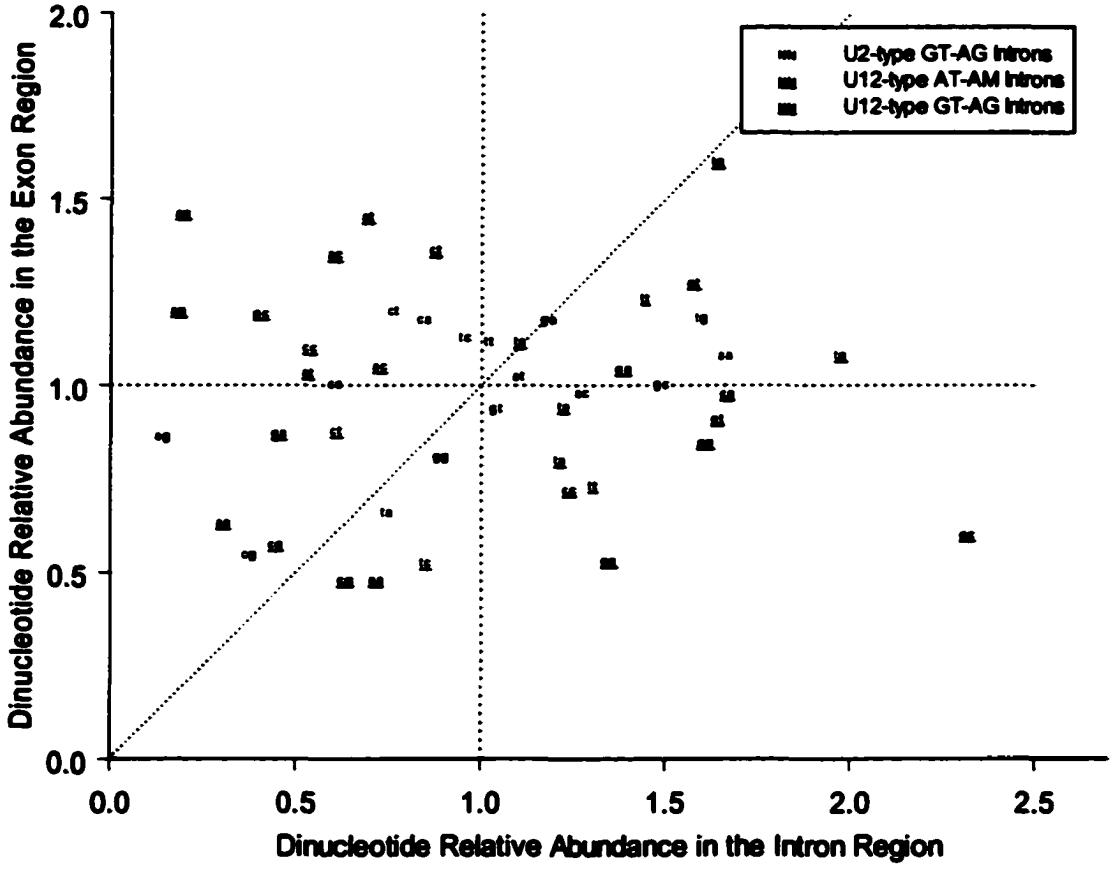

U2-type GT-AG Introns
U12-type AT-AM Introns
U12-type GT-AG Introns
Dinucleotide Relative Abundance in the Exon Region
Dinucleotide Relative Abundance in the Intron Region
0.0
0.5
1.0
1.5
2.0
2.5

Figure 5

Table 1. The list of U12-type introns.

GeneID	Location			Termini	Donor Site	Branch Site	Acceptor Site	DistBA
	Chr	Start	End					
At1g02750	1	603679	603774	at_ac	aag|atatcctt	ttccttgac	aatgtgatttatttgaaattccttgAccgataattttac|agtgagcaga	14
At1g04130	1	1074425	1074526	at_ac	atg|atatcctt	ctccttagc	tgttcaatgcgctttatgtccctccttAgcgctctaaac|atgttttcag	12
At1g06890	1	2112704	2112599	gt_ag	att|gtatccta	tatcttaac	aaccaattcgatgctgttatcttaAccaattgtacacag|atgacaaata	15
At1g11110	1	3710856	3710951	gt_ag	agc|gtatcctc	ttccttaac	ttatctggtctctgttatttccttaAcaagaaacagcag|aagtaggtgt	14
At1g11890	1	4011696	4011982	gt_ag	cca|gtatcctt	ttccttgac	agaccagatatgtatgcttccttgAccagaaaactttag|ttacatcata	15
At1g17145	1	5861748	5861660	gt_ag	tca|gtatcctt	ctccttatc	gaaaatgtctacttttgagtcctccttAtcatatgtaag|atgccctttt	12
At1g18090	1	6226767	6226945	gt_ag	acc|gtatcctt	ctctttgac	tttgcagttctgttctctttgAcacattaaacttatcag|agatctgtcg	18
At1g23900	1	8442060	8442234	gt_ag	aca|gtatcctt	aaccttaag	aaatagatttgttcctgtttaaccttaAgtcacgtgcag|agatcttaat	12
At1g24050	1	8507200	8507305	at_ac	ttg|atatcctt	atccttgac	cctgaattgtgttatccttgAcgagttattggttattac|aagagggtac	19
At1g24706	1	8755013	8755162	gt_ag	ccc|gtatcctt	ttccttgac	ccaaatccaagaatagcaatttccttgAcattgctctag|aatctatgac	12
At1g26660	1	9212563	9212827	gt_ag	act|gtatcctt	ttccttgac	gaaatattcttgtttccttgAcggttactcccactccag|ctctgacttg	19
At1g29630	1	10351407	10351491	at_ac	ccc|atatcctt	tttcttaac	aattaattaacatctaattgatttcttaAcataattcac|atcgatgcca	11
At1g29940	1	10486473	10486378	at_ac	gca|atatcctt	ttccttgac	tgtttgtgttgtgtctgatttccttgAcaagtatgatac|tttggttaga	13
At1g31660	1	11331328	11331222	at_ac	att|atatcctt	ttccttgac	gttatgccatgttatgttatttccttgAcaacaagtcac|ttgccccagt	12
At1g32400	1	11690408	11690243	gt_ag	tgc|atatccta	ttccttaat	ctaaaagtcttaacctttcatttccttaAttgagaatag|tactctcttc	11
At1g48050	1	17324178	17324251	gt_ag	act|gtatcctg	ttctttaac	gtttcgctaccatttttcttggttctttaActggcttcag|ttcttgatgc	11
At1g49160	1	17781408	17781268	gt_ag	agt|gtatccct	ttctttaac	aattcattatttctttaActctcttgaacaaataatcag|ttttaaggga	22
At1g50510	1	18318618	18318785	gt_ag	caa|gtatcctt	ttcctttac	cagcattgtaaaaccttgagttcctttAcacatttccag|acattgcgct	12
[redacted]	1	19895039	19894959	at_ac	tta|atatcctt	aatcttaac	atggccaaaaatgtattgtaaaatcttaActcagaacac|ctctttaagt	11
[redacted]	1	19896456	19896355	at_ac	att|atatcctt	ttccttgac	tggatgtcattttccttgAcatcattcatttgtttgaac|ccagtcaggt	21
At1g54460	1	19942066	19941998	gt_ag	ctc|gtatcctt	atccttagc	tctaaccactaaaggacttttatccttAgccatatatag|atcagctaca	12
At1g56280	1	20673690	20673593	at_ac	aag|atatcctt	ttccttaat	gaatcacattcctctgttgaattccttaAtgagttttac|ttgcagcgaa	11
At1g60070	1	21748447	21748332	gt_ag	gca|gtatcctt	aaccttgac	ttttaagttcatcttgtgagaaccttgActtacgtctag|agatctgaat	12
At1g61150	1	22143060	22143160	gt_ag	aac|gtatcctt	ttccttatc	gcttatctactccttttgaacttccttAtcgagaaacag|cagagataga	12
At1g61210	1	22164603	22164894	gt_ag	tac|gtatcctt	taacttaac	tgggtttgtttataacttaActtgattttggaaacttag|aggagttttt	20
At1g67960	1	25081843	25081760	at_ac	caa|atatcctt	gtccttaac	ccaacatctttattctgttgtgtccttaActgtcttcac|ttctccattc	11
At1g73350	1	27176027	27175925	at_ac	aaa|atatcctt	tttcttgac	ctagatgtttttttgtaatctttcttgAccatatatcac|aaagctgcaa	12
At1g76170	1	28186023	28185933	at_ac	cat|atatcctt	gaccttaac	tcaacagaagatattcgttagaccttaAcatagttacac|gtatgcttat	12

Table 1. (Continued)

GeneID	Location			Termini	Donor Site	Branch Site	Acceptor Site	DistBA
	Chr	Start	End					
At1g76940	1	28504478	28504399	gt_ag	ctc|gtatcctа	ttccttaac	ggaaactatttctctccactttccttaAccgatgtctag|atatatttcg	12
At1g78420	1	29106763	29106891	gt_ag	tca|gtatcctt	ctccttgac	gggtttttttccctttttgtgtcctccttgAcatttatgag|gtgccctttt	11
[redacted]	1	29555440	29555341	gt_ag	att|gtatcctc	ttccttaac	ttccttaActttctttttcatccttggtgacgaacctcag|ttatttatcg	32
[redacted]	1	29556223	29556137	at_ac	ctc|atatcctt	agccttaat	catacacatagacgagttgttagccttaAtgtaaagcac|ctattcaagt	11
[redacted]	1	29557508	29557382	at_ac	att|atatcctt	ttccttgac	aacaaaatatgttgttttccttgAcgtacttgtttgcac|ccagtcaggc	16
At1g80210	1	29766155	29766046	at_ac	gat|atatcctt	ttccttcac	caagagtgttttgtagtcttccttcActaatttcgacac|agaatgacga	14
At1g80500	1	29873046	29873186	at_ac	gcc|atatcctt	ttccttaat	cagttgtaaatatctgtttcttccttaAtttatcaacac|atacccgact	12
At2g20230	2	8675023	8675144	gt_ag	ttt|gtatcctt	tcccttgac	aatgggtttgttgcggtttttcccttgAcagttttaaag|tactctattc	12
At2g21880	2	9274381	9274252	gt_ag	tca|gtatcctt	aaccttgac	gatctgtgtttctgcggtaaccttgActtgcgacttaag|atatgtgtac	14
At2g25310	2	10726954	10726713	gt_ag	ttt|gtatcctt	gtccttgac	tcttacttttgcggtccttgActaattggtttgtaaag|tcatgaagtg	18
[redacted]	2	11193817	11193727	gt_ag	act|gtatcctt	tttcttaac	gtgattttcttaActtagattgtttatttt[redacted]accacag|gcagttatgg	27
			11193734	gt_ag			cgtagctgtgatttttcttaActtagattgtttatttt[redacted]|accacaggca	20
At2g26590	2	11262716	11262477	gt_ag	gat|gtatcctc	ctccttgac	cagttacgaagttttctccttgAccatacctatttatag|gatcaaattg	17
At2g27840	2	11811390	11811501	at_ac	act|atatcctt	atccttgat	ccttgAttttgttttgctctgctttttttctgatgctcac|ggatgaggag	34
At2g30470	2	12931888	12931800	gt_ag	aaa|gtatcctt	tatcttaac	gacttctggtgactcattgtttatcttaAcaaattctag|tttgaatttg	11
NA	2	13119502	13119751	gt_ag	tca|gtatcctt	ttccttaac	gctcttgccatttttccttaActctgtttttttctttcag|attctttagt	20
At2g36010	2	15071167	15071261	gt_ag	act|gtatcctt	atccttaac	cataagatataaatacgcaatccttaActctagcaaaag|ctggtatcga	13
At2g36810	2	15378136	15378019	gt_ag	gct|gtatcctt	taccttaat	tgctcttggtgtattactgcttaccttaAtcacatgaag|atatctgcac	11
At2g39070	2	16259192	16259086	at_ac	tac|atatcctt	ctcctgaac	gaagatttgagcattatgtttctcctgaActtttatac|acagaaacta	11
At2g39810	2	16562023	16562307	gt_ag	att|gtatcctt	ctccttgac	ggatgcatagttgacatctccttgAcagaaaattaatag|atatcaccaa	15
At2g39960	2	16631631	16631373	at_ac	att|atatcctt	gtccttgac	aattctgaagaagatgtttggtccttgActatatgacac|gtatgtagtg	12
At2g40835	2	16994801	16994903	at_ac	caa|atatcctt	gttcttaac	tatgctttctgcatgttgggttcttaAcgattctttcac|atatagatac	13
At2g41740	2	17365293	17365173	gt_ag	acc|gtatcctt	aaccttaac	ggactgtttgagcttgtgaaccttaActttttgttattag|agggactgaa	14
At2g42245	2	17547304	17547394	gt_ag	gca|gtatcctt	tgacttaac	cacgctttccgaaaaactgacttaAcaagcaatattag|atatgctaga	14
At2g43210	2	17909956	17910131	gt_ag	tat|gtatcctt	gatctttac	acgtgtgaatgcaatagggatctttAcatcagtttctag|acccatattc	14
At2g44150	2	18207554	18207643	gt_ag	gca|gtatcctt	ggccttgac	gctttgatcaagatttgttttggccttgActtgttaaag|atatatactt	11
At2g44270	2	18247293	18247211	at_ac	cat|atatcctt	taccttaac	gtcaacaaaagatattgtgataccttaAcgtagctttac|gtatgcttat	12
At2g44680	2	18375661	18375494	gt_ag	gca|gtatcctt	ttccttaat	ccacttgtcttagtgttctttccttaAtgcttattgcag|acattgatgg	13
At2g45240	2	18606938	18607027	gt_ag	caa|gtatcctt	gtccttaac	agtctattccagtgtaaacgtccttaAcctcaaaatcag|gaaacaaagc	13

Table 1. (Continued)

GeneID	Location			Termini	Donor Site	Branch Site	Acceptor Site	DistBA
	Chr	Start	End					
At2g47650	2	19489197	19489088	at_ac	atc|atatcctt	gtccttaac	cttgcttgctcttgtccttaActctgtgtgtgtggatac|aagacgaatg	19
At3g03340	3	786934	787252	gt_ag	aga|gtatcctt	tgctttaac	ttctctttccactcctgtgttgctttaActcgcccttag|ataccgagaa	12
At3g04630	3	1259668	1259776	gt_ag	ttc|gtatcctt	ttccttaac	acagaatcgatattgtttccttaAcgttatttaccacag|cgttgcaact	16
At3g05480	3	1587748	1587992	gt_ag	gtc|gtatcctt	atctttgac	ctttctatatatgctttgttatctttgActctatttcag|ccaacgtccc	12
At3g05700	3	1683177	1683077	at_ac	aag|atatcctt	gttcttgac	atagtctgttattaagttgtgttcttgAcaaaatattac|atgcaccgca	12
At3g06760	3	2133770	2133906	at_ac	aag|atatcctt	ttccttgat	tggtagaacttgtgtttattccttgAtgagaggttttac|gtacagcgaa	14
At3g06820	3	2152800	2152669	at_ac	gat|atatcctt	ttccttaac	gagtgttttgttgtcttccttaActaacctctctgacac|agaatgacga	17
At3g07100	3	2246959	2246871	gt_ag	gct|gtatcctt	atccttaac	cacatcatttagcttttagtaatccttaAcagaactaag|atataccgcc	11
[redacted]	3	4388199	4387910	gt_ag	atc|gtatcctt	ttccttgac	ctttgcttccttgActcaccttgtttgt[redacted]aagctacag|atctgttgca	26
			4387918	gt_ag			tgtgtaattctttgcttccttgActcaccttgtttgt[redacted]|aagctacaga	17
At3g16220	3	5498599	5498696	gt_ag	gtc|gtatccta	ctctttaac	tatgcaagtgcttctgaactctttaActctctagtttag|atgttatcat	14
At3g18750	3	6456511	6456321	gt_ag	tgt|gtatcctt	taccttaac	gcatctttttggtttgtaccttaActcaaggaaatcag|ttataaggca	15
At3g21070	3	7381345	7381258	gt_ag	ttc|gtatcctt	ttctttaac	attatctgttacaaaaactttctttaActccaatccag|acagcgaaca	12
At3g21215	3	7443078	7442982	at_ac	cca|atatcctt	cttcttaac	ttttttcttttctgtttgattcttcttaAcccaatgatac|atattcaaag	12
At3g24100	3	8703387	8703771	at_ac	ctc|atatcctt	ttctttaat	tctttaAttcttgcttttgttatgattgcgaatttgcac|gaggaagtca	33
At3g28370	3	10623358	10623507	at_aa	tag|atatcctt	ttccttgac	ccattctttcaaaagaatagtttccttgAcaaaaaacaa|gcagacttct	11
At3g44730	3	16296655	16296746	at_ac	aag|atatcctt	atccttgac	ttcctctgagaatcccgtaatccttgAccagtttctcac|attagacata	13
At3g46210	3	16985855	16985632	gt_ag	tct|gtatcctt	ttccttgat	cagtttttgctgctttttccttgAttcaattctgtttag|cttctaccat	16
At3g47990	3	17723846	17723550	gt_ag	tgt|gtatcctt	ttccttaat	tctccatttcgtttccttaAtggttattttgcgtcgtag|aatcattgta	20
At3g48260	3	17883035	17882932	gt_ag	agt|gtatcctt	ttctttgat	cagtcactaaagggtttctttgAtcagagagttcttcag|atacagagca	17
At3g49410	3	18334480	18334587	at_ac	aag|atatcctt	gtccttagc	agagctgtatttgcttggtctgtccttAgctttcaatac|agatcccaaa	12
NA	3	18783597	18783703	at_ac	tca|atatcctg	atccttaac	atgctcattgactattgaaaatccttaActtatgcgtac|tttagatctc	12
At3g51460	3	19102307	19102404	at_ac	gga|atatcctt	ttccttaac	tttatttgtaaaataaatctttccttaAccctgattcac|gctttaatag	12
[redacted]	3	19360482	19360367	gt_ag	gga|gtatcctt	tgccttgac	atgatctgcagtgcagaatgccttgAccaatccttccag|atattttgga	14
			19360395	gt_ag		ttccttcgc	cactttccttcgcattgagtgaggaagaatgatctgcag|tgcagaatgc	?
At3g53520	3	19850244	19850491	at_ac	atc|atatcctt	atccttaac	gatatgatgattctggttgatccttaAccatagttttac|aagacaaatg	13
At3g57410	3	21257746	21257584	gt_ag	acc|gtatcctt	aactttaac	ttactgctgtctgagatgttgaactttaAcatttcttag|aggtactgaa	11
At3g59310	3	21931164	21931056	gt_ag	ggt|gtatcctt	ggccttgac	atatttttcagctaaggccttgActtcacttgtccatag|aagtatactt	17
At3g59320	3	21933522	21933445	gt_ag	gat|gtatcctt	ttccctgac	cttaaacttctatgaacattttccctgActtgttcatag|aagcatattt	12

Table 1. (Continued)

GeneID	Location			Termini	Donor Site	Branch Site	Acceptor Site	DistBA
	Chr	Start	End					
At3g59340	3	21937790	21937699	gt_ag	ggt|gtatccttt	ttccttaac	tatgcacattctcgttgacgttccttaAcctgtctatag|agccatattc	12
At3g60250	3	22278963	22278836	gt_ag	gca|gtatccttt	ttccttaaa	tcattgtatcttagcgttttccttaAatgacttgtgcag|atattgatgg	14
At3g61710	3	22848274	22848171	gt_ag	tca|gtatccttt	atccttgac	gtatatgaagtttatccttgAccagtaaccccttttaag|atgtcaagtt	19
At3g62830	3	23241148	23241229	at_ac	atc|atatccttt	ttccttgac	aattgttgtttgtgtttgttccttgActttgattgatac|aagacgaatg	14
At4g00810	4	345885	345541	gt_ag	act|gtatccttt	atccttaac	aatgtttgggttaggtatccttaAccctattattgatag|tctgataaga	16
At4g01480	4	626473	626552	gt_ag	act|gtatccttt	atctttaac	accatttctttttgctatctttaAccaaaatataacag|acaagaagaa	15
At4g02200	4	972608	972757	gt_ag	aag|gtatccttt	ttccttgac	ttgtgttccttgAcgaatacttaccttgaacaaaagcag|agactttaca	27
At4g02480	4	1086553	1086414	gt_ag	tat|gtatccttt	tttcttaac	tcagtgtcttttggtaaaaatttcttaAccatcttctag|atatttcaac	12
At4g02560	4	1124966	1124732	gt_ag	cgt|gtatccttt	ttccttagc	atccttcacttcaaagttagtttccttAgctttcagcag|ataacaaaga	12
At4g03560	4	1583661	1583815	gt_ag	tga|gtatccttt	tttcttgac	ttttcagtataagatcgtttcttgAcgggaaaaactcag|ctacctttg	15
At4g04910	4	2494488	2494184	at_ac	ctg|atatccttt	gaacttaac	tcatttatttcgatttttgtgaacttaActggtgattac|tggacatgga	12
[redacted]	4	3161270	3161352	gt_ag	tct|gtatccttt	tttcttaat	tctctgttttagtttttttcttaAtgttacgccgatatag|atattgcgct	17
[redacted]	4	3162104	3162204	gt_ag	gca|gtatccttt	ttcctcaac	ctttatagaaagctgttgttttcctcaAccttttttgtag|ttcttacggg	12
[redacted]	4	5098265	5098368	gt_at	tca|gtatccttt	gtctttaat	attgattatgtgattatgtgaagtctttaAttggtga[redacted]|tttagatatg	10
			5098373	gt_ag			ttatgtgattatgtgaagtctttaAttggtga[redacted]tttag|atatgtgcat	15
At4g12790	4	6483467	6483325	gt_ag	gga|gtatccttt	agccttaac	atgatatgttttggattagccttaActtttaatccatag|gtatcttgag	15
At4g13345	4	6732628	6732712	gt_ag	ttt|gtatccttt	ttctttaac	atgccttcagagagatgttttctttaActatttctctag|ttgttctatt	13
NA	4	6889350	6889654	at_ac	ctc|atatccttt	ttctttaat	tttaAttcaagcttttgatttttgatgcgtaatcgtcac|gtggaagtca	35
At4g15030	4	7544863	7544970	gt_ag	aga|gtatccttt	ctccttgac	gtggataagtgcatgtctgttctccttgActgtttgaag|atactgtata	11
At4g15810	4	7952425	7952008	at_ac	aga|atatccttt	ctccttgac	tttgatctccttgActttgattttgatttgatgggtcac|ttgatgacat	26
At4g15950	4	8006562	8006658	gt_ag	tga|gtatccttt	aaacttaac	atttcttattgtgtgttttcaaacttaActgaatatcag|acctctgaaa	12
At4g17640	4	8791716	8791864	gt_ag	gaa|gtatccttt	ttctttaac	tctttggtttttgtcattctttaAcacacaaaaaaacag|acatagatgg	16
At4g17895	4	8905841	8905948	gt_ag	act|gtatccttt	atccttaag	ttatgagatgcattttgatttatccttaAgctgtcacag|atttttcacc	11
At4g19150	4	9437524	9437157	gt_ag	tcc|gtatccttt	gtccttaac	ttttttgtccttaAcatcatgatctcttggggaaatag|actacattta	25
At4g23330	4	11158191	11158320	gt_ag	gtt|gtatccttt	atccttaat	agaaatttgttctagaatccttaAtaaatcaatctgcag|aaagcacctg	16
At4g25290	4	11907099	11907018	gt_ag	gca|gtatccttt	taccttaat	atattgaagatagtgagattaccttaAtacacataatag|tctcgagacc	13
At4g25340	4	11927094	11926988	gt_ag	ttc|gtatccttt	aatcttaat	tttagaaatcttaAtgttttcattgcgactatattatag|agatggtatt	26
[redacted]	4	12763787	12763637	gt_ag	taa|gtatccttt	ccctttgac	agtccatgcgatacttgtctccctttgActcaataccag|atattttggt	12
[redacted]	4	12765180	12765095	gt_ag	acc|gtatccttt	ctccttaac	acaagaaaccttttttatgcctccttaActaacttataag|ctacctcgag	13

Table 1. (Continued)

GeneID	Location			Termini	Donor Site	Branch Site	Acceptor Site	DistBA
	Chr	Start	End					
At4g28770	4	13177930	13177849	gt_ag	ttc|gtatcctt	atccttgat	actattcagttgtgatgtttaatccttgAtaggtcttag|tattcaattc	11
At4g29330	4	13411669	13411762	gt_ag	ggt|gtatcctt	taccttgac	agtttttgttttgctgttgtaccttgAcattctcaacag|taataaaatc	13
At4g30160	4	13719584	13719683	gt_ag	agc|gtatcctt	ctccttgac	tgaaatatctacatgatgttctccttgActataagttag|tggtattgag	12
At4g30860	4	13991201	13991299	gt_ag	gca|gtatcctt	ttccttgac	ttcacaacattagtgtgtcttttccttgActgagttaag|atatttacct	11
At4g30900	4	14006234	14006348	at_ac	gct|atatcctt	ttctttgac	atatatcagtattttggttttctttgAcaagctgcctac|acattccaac	13
At4g36850	4	16318623	16318523	gt_ag	ttt|gtatcctt	taccttaat	ggttgagctttgctaccttaAttcgaaaaatcaaaacag|ataatactgc	19
At4g37210	4	16478622	16478706	gt_ag	att|gtatcctt	tatcttaac	attcagttgttggattatatcttaActtgaacgtttcag|aaacttccgc	15
At4g38240	4	16898080	16897966	at_ac	ggc|atatcctt	tttcttaac	agtttgacgtttctaagacttttcttaAcaaatatccac|ttactgggat	12
At5g03740	5	982698	982810	at_ac	tcc|atatcctt	ttccttgac	tgctttactctgtttttgacttccttgActctatctcac|cgagcctgag	12
At5g06590	5	2017335	2017517	gt_ag	tct|gtatcctt	ttccttaat	aatgttgttctggatgatttttccttaAtgttttcttag|tctagataat	12
At5g06620	5	2035149	2035245	gt_ag	ctt|gtatcctt	gttcttaac	gttatctctttgctatttacagttcttaAcatccagaag|ttctgacaaa	11
At5g08390	5	2701477	2701751	gt_ag	tcc|gtatcctt	agccttgac	gtattgcttagccttgActtatgtgatacattttataag|aggaatttgt	23
At5g08430	5	2718636	2718762	at_ac	aga|atatcctt	ttccttgat	ctttgtttataagttttccttgAtttcatgagttgtaac|gctgtccgag	17
At5g08500	5	2749580	2749947	gt_ag	ctt|gtatcctt	ttcctttac	ggtccatatctgctctttcctttActtaaatattggcag|acttactggt	16
At5g08630	5	2802026	2801798	gt_ag	tca|gtatcctt	ttgcttaac	tgccgtgtcagaactatttgtttgcttaAcataaaatag|attgtacttg	11
At5g09820	5	3056256	3056170	gt_ag	ttc|gtatcctt	ttccttgat	ttatggattggattccttgAttctgagaaattattgcag|atacttagac	20
At5g09920	5	3096831	3096978	gt_ag	aga|gtatcctt	ttccttaac	tttgtgagaatatgctattccttaActcaacttttgcag|aatactaagc	15
At5g10060	5	3147974	3147855	gt_ag	aaa|gtatcctc	tttcttgac	tttttcgatgtttggtttcttgActctttttggaatcaag|ctctgtcaca	17
At5g11580	5	3719580	3719685	gt_ag	ttt|gtatcctt	atccttgac	actctccattggttgcttcaatccttgActggtaaatag|atggccaact	12
At5g13570	5	4367881	4368077	gt_ag	tct|gtatcctt	ttccttgac	ttgttttctgtggatatttttccttgAcgtggcctatag|tattcaacag	13
At5g14850	5	4804104	4803795	gt_ag	tct|gtatcctt	ttccttaac	gcattttattctggattgcttccttaActtttttagtag|ctcttttgtc	13
At5g17440	5	5750325	5750450	gt_ag	aga|gtatcctt	tgctttaac	atttcgctcctctactgttttgctttaActcgtctttag|atataaagat	12
At5g20520	5	6945760	6945631	gt_ag	gaa|gtatcctt	gaccttgac	atactttgtcttgttacgaccttgAccttagaatcccag|acatcgctca	15
At5g22220	5	7341609	7341724	gt_ag	aat|gtatcctt	ctccttgac	tgatatctgtttagagctccttgActctttcatcattag|ctggaccaga	16
At5g22370	5	7384778	7384852	gt_ag	acc|gtatcctt	tttcttgac	aaattgcaggtgtatgttgtttcttgAcattgattttag|ttatgagtgt	13
At5g22650	5	7512510	7512639	at_ac	gca|atatcctt	ttccttaat	ggttgtgtattccttaAttcttgttgattgagtttccac|ggatgacttc	23
At5g23575	5	7925927	7926225	gt_ag	ctc|gtatcctt	gtacttaac	caatttgttatatgttggtacttaActttagatttgcag|acttactggt	15
At5g24450	5	8328174	8328291	at_ac	aag|atatcctt	gtccttagc	tcattttggatatctttgatcagtccttAgcattcatac|agatcccaaa	11
At5g25270	5	8739260	8739076	gt_ag	atc|gtatcctt	ttccttaat	tccttttaccggtaaaattccttaAtggttccatcttag|atgttgaaga	15

Table 1. (Continued)

GeneID	Location			Termini	Donor Site	Branch Site	Acceptor Site	DistBA
	Chr	Start	End					
At5g26180	5	9128913	9129076	gt_ag	cca|gtatccttt	tttcttaat	tgtccgtctatgagggcttttcttaAtgccaatttgtag|atatcgaagt	14
At5g26990	5	9451621	9451940	at_ac	aag|atatccttt	ttccttaac	ttttatctctaacagttacattccttaAcaaaatattac|gtgacgcgga	12
At5g27000	5	9459889	9459981	at_ac	aag|atatccttt	ttccttaac	agatacttaggttttgtttgttccttaActcttgattac|gttagagatc	12
At5g27380	5	9628389	9628282	at_ac	aaa|atatcctc	ttccttagc	cacaaaaaatggtgttgttcttccttAgcttcagaacac|gcataatata	13
At5g38380	5	14955401	14955507	at_ac	ttt|atatccttt	taacttaac	ttttaatggtgtggatagtttaacttaAcaacaagctac|caacaaaaga	12
At5g44200	5	17411251	17411118	at_ac	ctg|atatccttt	ttccttaac	ccgccaatccatgtgtttttttccttaActcattgattac|ctagaggtgg	13
At5g45760	5	18170205	18170300	gt_ag	agt|gtatccttt	caccttaat	aaaaatgtctcaaatcaacaccttaAtgagacatataag|atatgtagtt	14
At5g46740	5	18575775	18575683	gt_ag	ctt|gtatccttt	atccttaac	gtttctttatatgaattggttatccttaAccaatcacag|atttttcatc	11
At5g48790	5	19388631	19388498	gt_ag	gtt|gtatccttt	ttccttaaa	agagactcttgttttccttaAaagaagacaaaacaatag|actatccaaa	19
At5g49230	5	19567732	19567642	at_ac	aag|atatccttt	ttccttgac	agagtcattttgtgtattccttgAcgtgagagattttac|gtgcagcgaa	16
At5g49540	5	19713255	19713003	gt_ag	gac|gtatccttt	ttccttaac	ctctgtttgtctcagtaggtttccttaAcaatggagtag|atttgcgtat	12
At5g55130	5	21983013	21982903	gt_ag	cca|gtatccttt	ttccgtaac	agcattgaaaggctgtgtttccgtaActgtataaattag|atatgatatc	14
[redacted]	5	22765796	22765586	gt_ag	ttt|gtatccttt	caccttgac	atactatattgttctttaccaccttgAcaaatctaacag|attttgttaa	13
[redacted]	5	22767886	22767797	gt_ag	gtt|gtatccttt	ttccttgac	acgactttagaaatattatttttccttgAccctgtactag|atgttgcctt	12
At5g58100	5	23122393	23122519	gt_ag	cga|gtatccttt	atccttgat	tttatttgctagtgttaatatccttgAttcagaaacaag|ataaaaagac	13
At5g63700	5	25102356	25102262	at_ac	aga|atatccttt	caccttaac	gttacatttgaaacaccttaActctatccatttgtttac|attagaggag	19
At5g65180	5	25653869	25654054	gt_ag	aaa|gtatccttt	gtccttgac	gattgtttggtccttgAccatttgttttgttgggataag|ctttgtcaca	23
At5g66020	5	26011715	26011635	at_ac	gga|atatccttt	ttccttaac	gatcatcacttcagacaattttccttaAccctgcatcac|gttttcatag	12

The information for all of the U12-type introns identified in this study is listed in the order of their genomic location, including the gene identifier (GeneID), genomic location, the termini dinucleotides, the donor site sequence ([-3, 9] relative to the donor site, where "|' denotes the exon-intron junction), the branch site (the position is labeled from 1 to 9) and the acceptor site ([-40,+10] relative to the acceptor site, where "|' denotes the intron-exon junction), and the distance between the putative branch site and the acceptor site (DistBA). The branchpoint sequences have the consensus sequence TTTCTTAAC, and the position 8 is chosen as the presumptive branch site if the adenosine is in that position or the position 7 is selected otherwise. If neither of them is adenosine, a question mark will be put in the last column (DistBA). Only

one case is found in the gene At3g52180 (highlighted by the underline, see the text for details). The assignment of the branch site may not be accurate in the case of two consecutive adenosines in the positions 8 and 9 (Mcconnell et al., 2002), but this should not have a significant effect on the analysis of the distribution of the distances between the branch site and the corresponding acceptor site. Of the total 162 introns listed in this table, 4 introns (including one GT-AT introns and three GT-AG introns) share the common donor site with other 4 GT-AG introns, in the genes At2g26430, At3g13460, At3g52180 and At4g09720 (the gene identifiers are shaded in the table), respectively. 50 AT-AC introns, 1 AT-AA intron, and 107 GT-AG introns are contained in the non-redundant Arabidopsis U12-type introns (alternative splicing transcript isoforms are excluded). Note that 4 genes (that is, At1g54370, At1g79610, At4g07390, At4g27640 and At5g57160) have multiple distinct U12-type introns in different locations.

Table 2. The fate of U12-type introns after large-scale segmental duplications in the *Arabidopsis* genome

Block ID	Age	Gene1	Gene2	Ks
0102031203980	recent	At1g06890		0.7299
0103319703610	recent	At1g49160	At3g18750	1.205
0104000102440	recent		At4g02480	0.5794
		At1g01100	At4g00810	3.7616
			At4g02200	1.122
		At1g01060	At4g01480	0.7543
0203257711080	recent			0.8186
		At2g44680	At3g60250	0.7847
		At2g41740	At3g57410	0.5008
		At2g44150	At3g59060	0.9011
0204107902160	recent	At2g20230	At4g28770	1.0095
0204153002470	recent	At2g25310	At4g32130	0.6263
0204341201650	old	At2g46860	At4g01480	1.8172
0305000103160	recent	At3g03340	At5g17440	0.6363
0305033201380	recent			0.9514
0305052001580	recent			0.9555
0305290000580	recent			1.0943
0305328300280	recent			0.6644
0305331801560	recent			0.6112
0405128403640	recent	At4g17640	At5g47080	0.8517
0405237901400	old	At4g30160	At5g57320	1.6217
0505065400320	recent	At5g06390		0.5045
				0.9469
0505069001350	recent	At5g10080	At5g65180	0.9116

The list of gene pairs (Gene1 and Gene2) and the linked synonymous substitution rates (Ks) were subtracted from the recent analysis of the *Arabidopsis* gene duplications (Blanc et al., 2003; downloaded from http://wolfe.gen.tcd.ie/athal/dup). Those gene duplications might have arisen from large scale segmental duplications in the *Arabidopsis* genome in different ages ("recent" or "old"; see Blanc et al., 2003). The genes containing the U12-type introns (Table 1) are highlighted in green and yellow, for AT-AC and GT-AG termini, respectively. We manually checked introns paired with the U12 introns in the paralogs. 3 novel U12-type AT-AC introns (added to Table 1), 3 novel U12-type GT-AG introns and 2 U2-type GT-AG were determined based upon the non-cognate transcript spliced alignments and the U12-type prediction scores, and are highlighted in dark green, blue and white, respectively. 5 GT-AG introns with "weak" U12-type splice signals are indicated in gray.

CHAPTER 7. GENERAL CONCLUSIONS

General Discussion

Features and problems of GeneSeqer

The GeneSeqer program is distinguished from other spliced alignment programs in its accuracy, which stems from its splice site prediction model. GeneSeqer originally had a logitlinear model incorporated with GC contrast between introns and flanking exons (Kleffe et al., 1996; Brendel and Kleffe, 1998), with respect to the feature of low GC content in plant introns (Goodall and Filipowicz, 1989). This model relies on the plant-specific characteristic, therefore the earlier version of GeneSeqer is only applicable for plants. In the current version of GeneSeqer, the splice site prediction is implemented by Markov models (Zhang and Marr, 1993; Salzberg, 1997), which are applicable to a variety of species including human, mouse, Arabidopsis, maize, and yeast. The sophisticated splice site prediction enables GeneSeqer to accurately identify the exon-intron boundaries even in the instances of low sequence similarity or a very short exon (Haas et al., 2002). In contrast, most other spliced alignment programs only check the dinucleotide termini of the presumptive intron (Gelfand et al., 1996; Huang et al., 1997; Mott, 1997; Florea et al., 1998). This causes the sequence similarity to be overvalued by those spliced alignment programs, which is why they cannot identify mini-exons or accommodate low similarity. On the other hand, GeneSeqer does not overweight splice site prediction scores, thus it is still able to identify non-canonical splice sites on the basis of sequence similarity. Overall, the two functional components, splice site prediction and sequence similarity, are closely integrated by dynamic programming so as to generate the optimal spliced alignment (Usuka and Brendel, 2000; Usuka et al., 2000). Furthermore, the genomic localization is built in GeneSeqer based on the suffix array algorithm (Manber and Myers, 1993). This makes GeneSeqer not only fast but also easy to manipulate. In addition, the current version of GeneSeqer can also run in parallel computing mode on clusters.

However, the massive volume of data is still a challenge for GeneSeqer. Mapping 176,915 *Arabidopsis* ESTs on the *Arabidopsis thaliana* genome (released on Aug. 20, 2002; 117,276,964 bp in total) with GeneSeqer produced 355,349 alignments (including non-cognate spliced alignments) in 120 hours on a 1 GHz Pentium Pro III processor CPU. It takes an average of 0.8 seconds for

GeneSeqer to make one EST spliced alignment. Ignoring the genome size difference and paralogous loci, it would take at least 46 days to generate 5 million cognate alignments for human ESTs. In reality, the time for each alignment is much longer because introns in the human genome are much larger than those in *Arabidopsis*. This problem becomes even more serious when aligning full-length cDNAs.

User interface of GeneSeqer

GeneSeqer is a C program with standard command line arguments and simple text/html output. For the convenience of most users, several web sites (accessible at http://bioinformatics.iastate.edu/cgi-bin/gs.cgi and http://gizmo1.zool.iastate.edu/cgi-bin/PlantGDB/GeneSeqer/PlantGDBgs.cgi) were established with the pre-processed EST sets for plants, *Drosophila* and *C. elegans*. The online service also includes an elaborate image map to indicate the locations of alignments, as well as the putative gene structures and open reading frames.

The JAVA program MyGV was also developed for local users to interactively browse the GeneSeqer output. MyGV can also load GenBank annotations and gene predictions to compare with spliced alignments from GeneSeqer. Additionally, MyGV is capable of running external programs such as GENSCAN (Burge and Karlin, 1997), or remote web service like GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi), and display the results directly on the MyGV panel.

For some species, such as *Arabidopsis*, the genome and the transcript sequence data are relatively stable. Hence, it is very useful and efficient to share the spliced alignment data with the community after creating the final map of the cDNA/EST to the genome. To demonstrate this, AtGDB (*Arabidopsis thaliana* Genome DataBase, accessible at http://www.plantgdb.org/AtGDB/) was established at Iowa State University, allowing users to easily query, browse and analyze the spliced alignment data of their interest without needing to learn how to use GeneSeqer.

Application of spliced alignment in *Arabidopsis thaliana*

We did not restrict our work to the development of bioinformatic tools, but also applied the GeneSeqer program to the solution of practical biological problems. One of these efforts was to improve the *A. thaliana* genome annotation using *Arabidopsis* transcript sequence data as mentioned above. *A. thaliana* is the first plant genome completely sequenced (The *Arabidopsis* Genome Initiative, 2000), and the genome annotation is well maintained by TIGR (The Institute of Genome

Research). Despite this, our analysis still detected at least 1,000 incorrect gene structures in the recent *Arabidopsis* genome annotation.

The EST-confirmed data, in turn, can be also utilized to further improve splice site prediction and gene prediction. For instance, more than 400 GC-AG introns confirmed by ESTs in the previous study provide a good training data set for the prediction of GC-AG introns. We also made a detailed analysis of the U12-type introns in the *Arabidopsis* genome. The results suggest that the sequence context of U12-dependent 3' splice sites (3'ss) may play an important role in the 3'ss selection in addition to the space constraint. This implies that in addition to considering the conserved donor site and branch site signals, the *ab initio* prediction of U12-type introns should include acceptor site signals even though their information content is weak. Interestingly, there is a small portion of GT-AG introns with low prediction scores for their donor sites or/and their acceptor sites. That is, those introns do not have U2-type splice signals typical of most GT-AG introns, which include some but not all U12-type GT-AG introns. We may term the introns other than U12-type GT-AG introns as "weak" U2-dependent GT-AG introns or "weak" introns. It will be interesting to find out whether the "weak" introns are spliced efficiently *in vivo* or whether the splicing of "weak" introns requires the involvement of some specific trans-acting elements. The collection and the analysis of the "weak" GT-AG introns may reveal features and potential biological roles of the "weak" introns and improve the prediction of those uncommon splice signals.

Recommendations for Future Research

Improve the performance of GeneSeqer

Full dynamic programming is the main reason why GeneSeqer cannot handle the human genome or other long gene structures, and is unnecessary when the target sequences have near-perfect matches with the exon sequences in the genomic DNA. To reduce the computing effort but keep the accuracy, we can use a blast-like approach (Altschul et al., 1997; Kent, 2002) to identify the high-scoring segment pairs (HSPs), and then only apply the GeneSeqer dynamic programming algorithm to fill in the dangling region between the neighboring HSPs.

Comparative genomics and spliced alignments

The comparison of genomic DNA sequences is becoming more important with a growing number of genomic sequences from different species available. Correspondingly, many bioinformatic

techniques have emerged to address this issue in the last few years (Miller, 2001; Mathe et al., 2002). Some of them aim to identify genes based upon the assumption that protein-coding regions are more conserved than other flanking regions in the course of evolution and thereby the structure of homologous genes is conserved. Nevertheless, intron gain/loss and conserved non-coding sequences (Mural et al., 2002) contradict this assumption and cause difficulties for the techniques based upon it. Alternatively, we could compare genomic sequences mediated by spliced alignments in order to get a better understanding of the evolution of gene structures and then devise an appropriate strategy.

Knowledge-based gene prediction

Spliced alignment is one important method for the identification of gene structures based on sequence similarity and splice site prediction. EST is the major data source, however, EST spliced alignments may usually only reveal partial gene structures. Therefore, as discussed in the first chapter, one solution is to combine spliced alignments with *ab initio* gene prediction. A tentative algorithm is described as follows. First, generate the "knowledge" sequence for the corresponding genomic sequence according to GeneSeqer output. Then, establish rules that specify how to utilize the knowledge sequence. For example, start codon and stop codon cannot occur in the confirmed intron regions, intron cannot be predicted to overlap with confirmed internal exon region, and so on. Based on the knowledge-rule, we could make knowledge-based gene predictions. In practical implementation, the rule is a conditional probability of the prediction based on the available knowledge, that is, a value between 0 and 1. This eliminates a large number of incompatible states (that is, the conditional probability is zero), and seeks the optimal prediction among compatible solutions. On the other hand, the knowledge-based gene prediction will behave as the *ab initio* gene identification if there is no knowledge available. Certainly, the knowledge does not necessarily have to come only from spliced alignments or sequence similarity. It may improve the gene identification to adopt helpful information whenever possible, which is beyond the ability of most gene prediction programs to date. To address this issue, knowledge-based gene prediction can be designed to accept multiple knowledge sequences in different forms from different resources with specific corresponding rules. For example, one piece of valuable information may come from RepeatMasker (Smit, AFA & Green, P.; RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) to mark the repetitive or low complex regions, which typically contain few genes. PromoterInspector (Scherf et al., 2000) or other promoter prediction programs can be another knowledge resource to provide a better prediction for the diverse promoter rather than the simple TATA position weight matrix exploited in GENSCAN (Burge and Karlin, 1997).

This design also brings another advantage. Users can interactively change the knowledge and the related rules and generate a more reasonable gene prediction in connection with available information. Therefore, knowledge-based gene prediction is not only used in the automated genome annotation, but is also utilized with a graphic user interface to allow human interaction.

Overall, almost all gene prediction programs attempt to make good gene predictions based on general gene features and limited information, which may be applicable to a large portion of genes but not all. Knowledge-based gene prediction can be applied to improve the prediction of genes pertaining to specific external knowledge.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-402.

Brendel, V. and Kleffe, J. (1998). Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. Nucleic Acids Res *26*, 4748-57.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res *8*, 967-74.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. (1996). Gene recognition via spliced sequence alignment. Proc Natl Acad Sci U S A *93*, 9061-6.

Goodall, G.J. and Filipowicz, W. (1989). The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell *58*, 473-83.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhar, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biology *3*, research 0029.1-0029.12

Huang, X., Adams, M.D., Zhou, H., and Kerlavage, A.R. (1997). A tool for analyzing and annotating genomic sequences. Genomics *46*, 37-45.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-64.

Kleffe, J., Hermann, K., Vahrson, W., Wittig, B., and Brendel, V. (1996). Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. Nucleic Acids Res *24*, 4709-18.

Manber, U. and Myers, E. (1993). Suffix Arrays: A New Method for On-Line String Searches. SIAM Journal on Computing *22*, 935-948.

Mathe, C., Sagot, M.F., Schiex, T., and Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res *30*, 4103-17.

Miller, W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics *17*, 391-7.

Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci *13*, 477-8.

Mural, R.J., et al (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science *296*, 1661-71.

Salzberg, S.L. (1997). A method for identifying splice sites and translational start sites in eukaryotic mRNA. Comput Appl Biosci *13*, 365-76.

Scherf, M., Klingenhoff, A., and Werner, T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol *297*, 599-606.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Usuka, J. and Brendel, V. (2000). Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J Mol Biol *297*, 1075-85.

Usuka, J., Zhu, W., and Brendel, V. (2000). Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics *16*, 203-11.

Zhang, M.Q. and Marr, T.G. (1993). A weight array method for splicing signal analysis. Comput Appl Biosci *9*, 499-509.

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I am deeply indebted to my supervisor Prof. Dr. Volker Brendel for kindly providing guidance throughout the development of this study. His comments have been of greatest help at all times. Gratitude also goes to my co-major professor Dr. Srinivas Aluru for his instruction and kindly help. I also extend my sincere thanks to my POS committee members Dr. Thomas Peterson, Dr. Patrick S. Schnable and Dr. Gavin J.P. Naylor for their guidance during my education.

Finally, I acknowledge all persons in Dr. Brendel's group, for their efforts during my educating and I also extend my thanks to Ms. Jacqueline E. Townsend for her kindly help in the preparation of this thesis.

Especially, I would like to give my special thanks to my wife Yun whose patient love enabled me to complete this work. This thesis is dedicated to my parents whose inspirations make me work hard.